

Human cloning requires a moratorium, not a ban

Unprecedented media coverage of cloning highlights a genuine need for reflection by society. Research into animal cloning need not be hindered, but a declared moratorium on human cloning is desirable.

The history of science suggests that efforts to block its development are misguided and futile. Yet, following the news that scientists in Scotland have successfully cloned a lamb from an udder cell taken from an adult sheep (*Nature* 385, 810–813; 1997), calls for an outright ban on human cloning have come from bodies and individuals as diverse as the Vatican, the US Biotechnology Industry Organization, and Joseph Rotblat, the 1995 winner of the Nobel peace prize. Politicians have not been deaf. Countries already banning the procedure have been smugly pointing this out. Other places — such as New York state — have seen hastily drafted bills seeking the same goal.

The reaction of the scientific community has been justifiable concern that such moves are in danger of throwing out the baby with the bathwater (see pages 8–9). Few people indeed would condone the idea of human cloning for political or aesthetic reasons, or even out of sheer curiosity. But there are purposes for which aspects of human cloning could technically be highly desirable — for example, in order to generate skin grafts for burn victims, or other 'spare-part' provision.

The arguments in favour of human cloning need to be heard, but now is also certainly a time for wide-ranging circumspection. True, recent developments may not represent an overnight invitation to the moral collapse of the human race that some critics predict. Furthermore, there are many technical reasons to believe that the cloning of adult humans will remain unfeasible for years to come. But there is no escaping the legitimate concerns arising from the fact that a lamb cutely named Dolly represents an irreversible development of breathtaking implications.

Circumspection requires the avoidance of both precipitous action based on panic, and inaction based on complacency. The challenge facing legislators, as in so many aspects of modern science in general, and modern genetics in particular, is to chart a way forward that balances two requirements. One is to maintain the maximum freedom of scientific enquiry, while recognizing that this is not necessarily an inalienable right. The second is to respect the commonly perceived dignity of human life, while accepting that, whatever some philosophers and religious leaders may say, this is not inevitably compromised by human cloning.

Experience in striking such a balance in recent years has been mixed, particularly in the United States. There, initial reaction to the challenge of recombinant-DNA technology was exemplary. Both the industrial promises and potential environmental threats of the new technology were recognized promptly. A brief initial moratorium, called in 1976 by scientists such as Paul Berg and colleagues to take stock of the situation, was soon followed by strict regulation. But this was steadily loosened as US society became increasingly confident of its ability to understand and handle the potential abuses and dangers involved.

In contrast, the US record on embryo research, a prey to the aggressive zealotry of the anti-abortion movement, reveals less to

be proud of. Where anti-abortionists have been able to wield their influence, in particular over the research agenda of the National Institutes of Health, they have been able to block all research in the field. But where their political ideology has discouraged interference, namely in the operations of the marketplace, they have sat on their hands, allowing biocommerce potentially to burgeon even in areas of public sensitivity. The calamitous and perverse result is that research for which it is forbidden to use public funds can now be carried out unregulated — and unmonitored — in private laboratories.

The experience with genetic engineering technology shows that an explicit moratorium can be productive. On the one hand it sends a public signal that ethical considerations are important and need to be taken into account, and allows that to happen. On the other, the fact that a moratorium is, by definition, likely to be lifted at some point indicates that further development can be anticipated.

The embryo research debate underlines a second message. While the bodies set up to regulate and monitor new developments in science must reflect a range of public opinion, the net must not be cast so wide that they are held hostage to unremitting dogmatism. An excellent model here is the all-party Select Committee on Science and Technology of Britain's House of Commons. The committee's report on human genetics, published in the summer of 1995, may have represented a compromise between the views of members of two opposing political parties. But it still had sufficient teeth to embarrass the government into setting up a broad-ranging genetics advisory panel, and the insurance industry into drawing up guidelines of good practice. Its proceedings were also a model of openness.

Perhaps the closest US model is the now-defunct Office of Technology Assessment (OTA). The inelegant posture that Congress finds itself in over embryo research, as well as the surprise with which news of the Scottish cloning experiments was greeted by the political community, highlights the political folly of eliminating the OTA. At the same time, the cloning news has shown up a limitation of the technology assessment approach: how blinkered it can be, despite the best of intentions. The chairman of one of the panels that conducted Britain's recent Technology Foresight exercise, which made no reference to the possibility of human cloning, admits candidly that this was not considered sufficiently seriously.

A moratorium on the cloning of animals would go well beyond what society needs in order to take stock. But a declared moratorium on human cloning is desirable, even though it carries with it a possibility that will worry those who wish to pursue such research: that legislators will consider the potential benefits but decide that the risks to society are too significant for it to be permitted at all. The history of technology suggests, however, that highly regulated human cloning will, after all, be found to be a tolerable way to proceed. □



Tritium leak at US reactor sparks crisis for neutron source users

[WASHINGTON] Brookhaven National Laboratory (BNL) on Long Island, New York, faces a year-long closure of one of its main research facilities — as well as a desperate battle to regain public confidence — after the discovery of major leaks of tritium in ground water near its High Flux Beam Reactor (HFBR).

The closure of the reactor has left hundreds of chemists, structural biologists, condensed-matter physicists and materials scientists scrambling for 'beam time' at other neutron sources in the United States and Europe, most already oversubscribed.

The leak is thought to have sprung from a 67,000-gallon, concrete-lined canal used to cool and store spent fuel from the reactor. Senior officials from the Department of Energy travelled to the laboratory last week and assured the public that the facility will remain closed until the leak is dealt with.

Horizontal wells will be drilled to confirm the source of the leak. If it is the canal, the department will drain it, ship the spent fuel to Savannah River, South Carolina, for storage, and line the canal with stainless steel before restarting the reactor, which has been closed since December for routine refuelling. The process will take at least nine months, officials say.

Under pressure from local politicians including Senator Alfonse D'Amato (Republican, New York) and Congressman Michael Forbes (Republican, New York), the energy department has already issued a report on the incident which sharply criticizes BNL management for failing to respond to early requests from the county

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Closed for business: the High Flux Beam Reactor at the Brookhaven National Laboratory.

government to monitor ground water close to the reactor.

Some researchers fear that the tritium leak could damage the case for overdue upgrades at BNL and other existing sources. "We hope this doesn't cause any collateral damage" to the case for upgrades, one neutron scientist says.

Tara O'Toole, assistant secretary for environment, safety and health at the Department of Energy, travelled to Brookhaven last week to meet managers and to address an eight-hour public meeting in Riverhead, also attended by D'Amato and Forbes. O'Toole says that the restart "may take a little longer" than a year, adding: "I'd be surprised if it takes much longer."

But she warns that lack of public confidence could be "a major roadblock" to reopening the reactor, and calls on users inside and outside the laboratory to help to per-

suade the public that the reactor is needed for research. "To some extent, it is in their hands," she says. "They have to take responsibility for educating the community" as well as listening to its concerns "in a respectful manner".

Brookhaven has sunk 500 wells since 1989 to monitor ground water, chiefly at the boundary of its site. But last August two wells were sunk near the HFBR, and abnormal levels of tritium found in them in December.

More test wells and sampling confirmed tritium at levels of up to 50,000 picocuries per litre by mid-January, when the laboratory announced the problem. Further sampling at the beginning of February showed levels as high as 651,000 picocuries per litre — 30 times the drinking-water limit set by the Environmental Protection Agency.

Samples collected so far suggest that the tritium is distributed underground in a plume 1,000 feet long and 100 feet in diameter. It is probably moving at less than a foot a day, according to Henry Bokuniewicz of Long Island Groundwater Institute at the State University of New York.

"The risk to water supplies is non-existent at this stage," he says, adding that the tritium — which has a half-life of 12 years — should decay before it reaches the local water supply wells four miles away.

Bob Casey, head of safety and environmental protection at Brookhaven, says that, although the need for a groundwater monitoring well was identified in 1994, it was not dug because there was no particular reason to give it priority.

The decision might have been different, Casey says, if he had known that the canal was liable to leaks. According to 1963 design calculations unearthed in the energy department's new report, its permeable concrete walls are liable to leak "3 to 8 gallons a day".

BNL officials now fear that their recent

British peers want clearer research goals

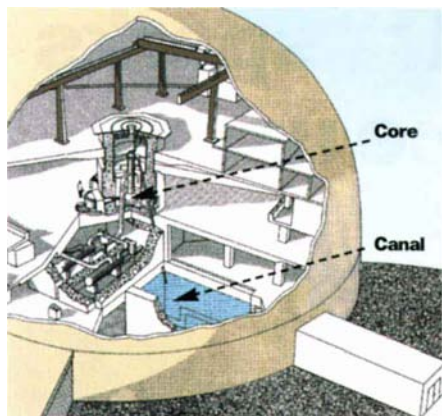
[LONDON] Britain's House of Lords has added its voice to those demanding a sharper focus to research activities funded through the European Commission in Brussels. It is seeking to have this expressed in plans now under discussion for the fifth Framework research programme (FP5), due to run from 1998 to 2001 (see *Nature* 385, 665; 1997).

In a report to be published in London today (6 March), the Lords science and technology committee says it is "astonished" at the breadth of the commission's current proposals for FP5. "We had been expecting the commission to come up with proposals that were much better focused than what has appeared so far," says Lord Selborne, the chairman of the Lords committee.

On the basis of a series of hearings held

over the past few months, which included witnesses from both Brussels and the British research community, the committee says it is unimpressed by the direct outcomes of research money spent through the commission. It suggests that similar results could have been achieved by spending the same amounts at national or global level.

Meanwhile, the commission has published a response to recent criticism. It says that the need to ensure that a five-year programme corresponds to the desires both researchers and "the people of Europe" means that its approval is inevitably a lengthy procedure, and insists that with a reduced number of planned programmes, "the new management structure will be lighter and more flexible". [SEE ALSO page 5]



The 'canal' or spent fuel pool, on the lower level of the HFBR, is the suspected source of the leak.

efforts to reassure Long Island residents of the laboratory's safety (see *Nature* 384, 205; 1996) have been undermined. People who had come to accept the laboratory's good faith now feel let down, Casey says. "There's a high degree of outrage," he says. "It will be quite an effort for us to recover."

John Axe, head of science programmes at the HFBR, says that his staff will now be "out scrambling with the rest of our users" looking for access to other neutron sources. He admits that the closure will put some of the facility's 300 users in "a terrible situation".

The most directly comparable facility, at the National Institute of Standards and Technology, can — like Brookhaven — offer both thermal and cold neutrons. "I'll do everything I can," promises Mike Rowe, head of that facility. "The problem is I am already full. I'll end up cutting everybody back so that no-one gets cut out."

Thermal neutrons are available at the High Flux Isotope Laboratory at Oak Ridge National Laboratory in Tennessee. "Neutron sources worldwide are oversubscribed," says Bill Appleton, associate director of Oak Ridge. "We'll just have to stretch everybody out. It's not a pretty situation, but one we'll try to make the best of."

Overseas researchers use HFBR too. Jeremy Bradshaw, a biophysicist at the University of Edinburgh, had been planning to spend three weeks there this summer studying how a feline leukaemia virus attaches itself to cell membranes. He says he cannot use the Institut Laue-Langevin in Grenoble, France, because less time has been allocated to British scientists there, and he may use a facility in Berlin. "I think I'll be able to continue with my work, but clearly at a reduced rate," he says.

BNL's director, Nick Samios, says he is confident that the leak will be plugged and the facility reopened. Of energy department criticism, he says: "They said we should have moved a little faster — but we've been working very fast on a lot of things." BNL's strategy, he says, was to study ground water at the boundary and at old industrial waste dumps on the site first.

Colin Macilwain

British universities face up to their day of reckoning

[LONDON] A mixture of delight and despair was to be found in British universities last week as each received official notification of how their recent research performance has been used to decide their core funding from central government next year.

Some institutions were celebrating the successful outcome of many hours of careful preparation of submissions for the latest four-yearly research assessment exercise, carried out by the funding councils which distribute money to universities in England, Scotland and Wales respectively (see *Nature* 385, 3; 1997).

University College London, for example, a high number of whose departments were awarded a 5 or a 5* in the exercise — indicating research at an international level of excellence — has been rewarded with a 9.5 per cent increase in its grant next year from the Higher Education Funding Council for England (HEFCE).

Much of this increase was the result of a growth of 23 per cent in the money allocated for the support of research. The University of Oxford, which performed best of all British universities in the exercise, with 75 per cent of its 'research active staff' considered to be in 5* departments, will similarly see its overall grant rise by 7.5 per cent.

In contrast, the University of Cambridge will receive only 4 per cent more. University officials say that this reflects the fact that, although a greater proportion of its research staff was submitted for assessment than Oxford, one result was to reduce the level of gradings achieved.

Some of the so-called new universities, created out of polytechnics in the early 1990s, have also done well. Nottingham Trent University — formerly Trent Polytechnic — which raised the research performance of many departments from a 2 to a 3a or 3b (on a scale of 1 to 5*), will have its funding for research increased by 74 per cent.

But the limited funding available inevitably means there have also been casualties. Some are large universities with established research reputations, in which a drop of one category in a large department — say from a 5 to a 4 — has been sufficient to trigger a substantial decline in overall research funding.

The University of Manchester, for example, has seen its allocation for research fall significantly — largely due to a fall in its ratings in some areas of medical research — and has ended up with a budget reduction of 0.1 per cent this year.

Manchester and other universities would have experienced even heavier cuts if it were not for a 'moderation factor' introduced to

mitigate the inevitable pain and disruption caused by a formula designed to reduce funding for poorly performing institutions.

The HEFCE announced that it was taking £12 million off the money that would have been allocated to top performing universities, and redistributing it to those at the bottom of the scale.

The result is that no university has seen its total grants reduced by more than 1.2 per cent. But it also means that some of the more successful institutions have seen their grants considerably reduced from what they would have been based on the funding formula alone — in the case of University College London, by £1.7 million.

Brian Fender, the director of HEFCE, defends the arrangement as a way of avoiding excessive financial stress on certain institutions. But he emphasizes that the safety-net procedure is likely to end next year, while those institutions that have required such assistance are being asked to draw up recovery plans.

The funding council's efforts at increased selectivity have been generally welcomed at a time when increasing student demand and level budgets are placing continuing pressure on higher education. "Spreading the misery evenly is not an option," says Gareth Roberts, vice-chancellor of the University of Sheffield and chairman of the Committee of Vice-Chancellors and Principals.

The Association of University Teachers (AUT) is calling for extra money for universities — unlikely under even a Labour government — and is also asking for more weight to be given to the regional roles of universities in assessing their grants.

At present, the main criterion used, in addition to student numbers, is the quality of research. But the AUT points out that among the losers this year have been several institutions, such as the University of Exeter, University of East Anglia, and University of Hull, that are the only research institution in their geographical area.

Another group of losers has been some specialist research centres, such as the School of Pharmacy in London. "Too little attention has been given to protecting regional research centres or specialist institutions across England," says David Triesman, AUT general secretary.

HEFCE is already drawing up suggestions for ways in which such factors might be taken into account in future. Also included may be consideration of the extent to which a university department has adopted priorities in line with the government's Technology Foresight exercise.

David Dickson

Heat rises over Brussels FP5 proposals

[BRUSSELS] Tensions are already high over the European Union (EU)'s fifth Framework research programme (FP5), due to run between 1998 and 2001, both among the various political bodies that must approve it and within the scientific community, concerned at the way it is likely to be carried out.

These tensions centre on how the European Commission intends to concentrate its proposed budget — expected to be ECU15 billion (US\$13 billion) — how the role of basic research will be protected, and how a proposed complex new system of management can be made to work efficiently.

Many of these issues surfaced last week at a meeting in Brussels of representatives of the programme's end-users, academic and industrial researchers, and advisory committees to the commission, called to discuss

the commission's draft proposal for the Framework programme. This will become a formal proposal by the end of March.

The draft proposal envisages three broad thematic programmes, interlaced with three policy-oriented 'horizontal' programmes (see *Nature* 385, 665; 1997). The thematic programmes include so-called key actions, research programmes directed to strategic goals, as well as untargeted programmes for generic technologies and basic research.

The meeting gave its general approval to the goal of concentrating funds on a limited number of activities. But those present pressed for more key actions — for example, in the areas of climate research and polar sciences — as well as an additional thematic programme for energy, which at present is subsumed within a thematic programme called 'Promoting competitive

and sustainable growth'.

Jan Borgmann, head of the European Science and Technology Assembly (ESTA), warned the meeting that the conflicts over where FP5 should concentrate its efforts are likely to increase when the commission's formal proposal is presented to the European parliament and the council of ministers. Everyone will lobby for their own interests, said Borgmann. "The commission will have a hard time," he said, but he stressed that ESTA will support its efforts to resist pressure to widen the scope of the programme.

Most participants at the meeting also approved the concept of the dual approach of targeted and untargeted programmes. But members of CREST, the EU's Scientific and Technological Research Committee which comprises representatives of member states, warned that the council of ministers is unlikely to be enthusiastic about the commission's concept of a separate budget for generic technologies and basic research, which would allow a 'bottom-up' approach to longer-term research aims. The council of ministers is made up of representatives of the member states, and the Framework programme requires unanimous approval.

Peter Idenburg, the Dutch representative on CREST, said that his committee had suggested that most, if not all, of the money should be concentrated in the thematic programmes on the key actions, and that these should have clearly defined output. Generic technologies and basic research would still be included in these categories, he said.

Rob Wright, British CREST representative, put it more plainly. "Governments will be extremely hard-nosed," he said. "They will not approve money for projects which are not palpably urgent."

But calls for reduced emphasis on non-targeted basic research were heavily criticized by Borgmann. "ESTA has had the feeling that the commission is now more friendly to basic research than it has been in the past", he said. "We would ask it not to deviate too far — even under pressure — from its desire to take a longer-term view of research than industry has to."

At the meeting, the commission defended the matrix-style management system that it has proposed for FP5. Although the proposed management structure has come under attack for its apparent complexity, the commission argues that it will increase flexibility. "It is the only system that could work," said Jorma Routti, director general of the research directorate. "As form follows function, in the philosophy of Frank Lloyd Wright, so structure will [follow objective in our management]." But commission officials were unable to say how the management structure will operate in practice.

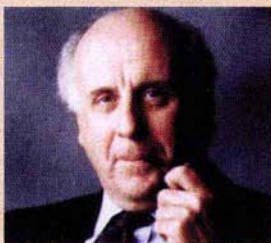
Alison Abbott

Lack of focus 'holds back European research'

[BRUSSELS] An independent five-year assessment of the European Union (EU)'s Framework programmes for research has concluded that European research is "underachieving" because of a lack of focus and a legal structure that hinders the development and implementation of a truly European strategy.

The report, by an independent panel of experts headed by Etienne Davignon, a former European research commissioner (right), stresses that the quality of research assessed was adequate but that the work often lacked clear goals. It calls for a radical rethink — a "leap forward" — in the formulation of strategy for the next, fifth, Framework programme (FP5), now being developed (see above).

Many recommendations are being implemented or are included in the European Commission's draft proposal for FP5. For example, procedural changes have been introduced to reduce what the Davignon report highlights as "unacceptable levels of bureaucracy and delays" in approving grants. But the report goes further, suggesting that the



commission should consider increasing staffing, or delegating some tasks to outside bodies.

In line with new commission thinking, it also says that 'structural funds' — the EU's subsidies to poor regions — should be used for research, and that FP5's budget should be more flexible, allowing priorities to be changed quickly.

But the report also makes a plea for two high-level changes to be made to the legal framework governing EU research, to allow programmes to run more effectively and efficiently. First, it says that the union's Maastricht Treaty should be changed to eliminate the requirement for a unanimous vote on Framework research programmes by the council of ministers. The power of single countries to veto an entire programme has meant that the EU research effort tends towards "an aggregate

of national and sectoral desire and ambitions". The report suggests that the treaty should be amended to allow for qualified majority voting.

Second, Davignon's panel says that the commission should be given a freer hand in implementing research programmes approved. Individual programmes are at present controlled by programme committees made up of officials from member states. The need for most decisions to be approved by these committees — 18 control the current Framework programme — slows down procedures, says the report.

The report recommends replacing these committees by a single, high-level committee. This "union committee", appointed by and reporting to the council of ministers, would monitor the commission's implementation of research programmes and act as an official advisory body to the council.

All the report's recommendations will be considered by the commission in preparing its formal proposal to the council and the European parliament at the end of this month. **A.A.**

AIDS drug row embroils health minister

[CAPE TOWN] South Africa's health minister, Nkosazana Zuma, an anti-AIDS campaigner who has enthusiastically supported a controversial AIDS drug, has come under fire for failing to distance herself from the criticisms of two separate bodies about unauthorized clinical trials of the drug.

The controversy over the effectiveness of the drug, and the conditions under which trials with it were carried out, threatens to develop into a serious conflict between South Africa's fledgling democratic government and the Medicines Control Council, the Gauteng Department of Health and the University of Pretoria.

Last month, the Medicines Control Council, the country's drug regulatory authority, suspended clinical trials of the drug, Virodene, which was developed by three researchers at the University of Pretoria's department of cardio-thoracic surgery. The drug's active ingredient is a highly toxic industrial solvent, dimethylformamide.

The researchers had proceeded with the trials without obtaining clearance from the council, and after their application to the university's ethics committee to conduct trials had been rejected (see *Nature* 385, 474; 1997). Normally applications have to be vetted by both bodies before trials can proceed.

The medicines council is concerned pri-

marily because the drug's active ingredient can cause irreversible and fatal liver damage, and has also been linked with the development of cancer in humans and experimental animals. The trials have been suspended until aspects of the drug's safety, weaknesses in the trial design, and the evaluation of results obtained so far, have been resolved. But several patients are claimed to be in remission as a result of receiving the drug.

Council chairman Peter Folb says that the clinical trials were designed in such a way that they could not be properly analysed statistically, and contained inadequate accounts of the patients treated. There had also been no validation of assays performed.

"The fact that the normal requirements for conducting a clinical trial were not met is unacceptable and potentially dangerous," says Folb. He adds that safety issues need to be resolved before any further work can be considered, and before patients who have already received Virodene are further exposed to the drug.

It is up to the University of Pretoria and the provincial health authority of Gauteng, which jointly employ two of the three researchers, to decide whether to initiate disciplinary procedures against them over the earlier trials.

A joint committee appointed by the two

authorities, which reported last week, has been even more critical of the researchers' infringements than the Medicines Control Council was. Chaired by the university's professor of genetics, Henk Huisamens, the committee found that the company Cryptopreservation Technologies, which has patented the drug, appears to have an unauthorized association with both the university and the researchers' department.

It found that the company's involvement "might have created a potential conflict of interest that would not have been to the advantage of open-ended research". One of the three researchers, perfusion technologist Olga Visser, was discharged from the department of cardio-thoracic surgery for medical reasons last July, and is now joint manager of this company, together with her husband.

The company has issued a statement saying that nothing will stop its research from going forward. The group says that a "top pharmaceutical research concern" will handle the next trial.

Zuma, meanwhile, has denied claims that she authorized the trials personally, saying she merely "encouraged" the three scientists to continue their work, of which she has been aware since last July. She claims not to have checked that they had followed the standard procedures as she had assumed they had done so.

She also appears reluctant to criticize the researchers' infringement of the conventional procedures intended to protect the scientific and ethical integrity of trials of new drugs. At a press conference last month she emphasized that side-effects of drugs had to be weighed against their benefits, adding that one patient had been on the drug for five months without displaying any apparent side-effects.

Zuma's passionate commitment to combating AIDS, which is now a serious problem in South Africa, stirred controversy last year when she approved R14 million (US\$3 million) funding for an educational play, *Sarafina*, which turned out to be a scam. The play was intended to raise awareness about the disease among the country's poorer communities, but it emerged that funding was being squandered on lavish payments to its producer and actors. A controversy developed both about the play's effectiveness in achieving its stated objective, and the manner in which its funding was approved.

The Virodene controversy has resulted in renewed calls for Zuma's resignation, both from the official opposition, the National Party, and the smaller Democratic Party. But President Nelson Mandela backed her strongly last year, praising her exceptional competence as a minister, and is likely to do so again.

Michael Cherry

'Data network threatens patient privacy'

[PARIS] France's data-protection authority and the council that regulates the medical profession have warned that a planned national computer network for the management of patient information threatens individual liberty and privacy, medical confidentiality and the professional independence of physicians.

Under a law passed last year, all physicians must equip themselves before 1999 with computers capable of transmitting patient information to the state health-insurance scheme. Although the main aim is to introduce savings and reduce the social-security deficit, the network will also improve diagnosis by allowing online consultation of databases.

But the availability of patient information also represents a goldmine for drugs companies interested in prescription practices. The sensitive data also risk being exploited by insurance companies, banks and employers.

In a statement issued in Paris two weeks ago, the Commission Nationale de l'Informatique et des Libertés (CNIL) — which oversees the 1978 data-protection act — says commercial exploitation of data should be banned. Both CNIL and the

medical council, the Conseil National de l'Ordre des Médecins (CNOM), also recommended that all data submitted to the network should be processed to ensure the anonymity of both patients and physicians. The CNIL said that connections between the network and the Internet should be banned.

Olivier Dubois, secretary general of CNOM, says that its recommendations — almost identical to those of CNIL — are aimed at protecting basic human rights while preserving researchers' access to data.

Under the recommendations, the use of computerized data would require contracts specifying the aim of a project in terms of public health or medical research. These would have to be approved by the local ethics boards established by the 1988 Huriot law governing clinical trials. Following approval, researchers would not need to seek the consent of individual patients, although patients would have the right to deny use of information about them.

Ethics boards would be authorized to waive the requirement of anonymity in cases where it was essential to the goals of the research, as in chronic diseases such as AIDS and cancer where there may be a need to track individual patients

Declan Butler

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Slipping behind: Russia's Mir craft is predecessor of its planned contribution to the space station.

NASA head predicts space station delay

[WASHINGTON] US officials expect construction of the international space station to begin at least six months after the original target date of November because of delays in Russian financing of a critical module. The 'service' module is intended to keep the station in orbit during the early stages of assembly and to house its initial crew.

Although the schedule slip had not been officially announced by Daniel Goldin, the head of NASA, by the beginning of this week, he has told a congressional committee that he anticipated a delay.

Goldin's reluctant admission came during a week of confusion and increased criticism of Russia's role in the project. Yuri Koptev, the head of the Russian space agency, announced that launch of the station's "first element" — on a Russian rocket — would slip until June 1998. But a surprised NASA official quickly dismissed that as premature.

Although Russian prime minister Viktor Chernomyrdin promised last month that his government would release money to Russian contractors building station hardware by the end of February, "it appears that's slipping into March," said a NASA spokesperson.

James Sensenbrenner (Republican, Wisconsin), chairman of the House of Representatives science committee, returned recently from a fact-finding trip to Russia frustrated at the lack of progress on the service module and the ever-receding deadlines. White House and NASA officials have been more patient, seeing Russian participation as part of a larger foreign policy initiative to engage that country's scientists and engineers.

Sensenbrenner and others fear political damage to the programme caused by delays and possible cost overruns should the Russians default. Marcia Smith, a space policy expert with the Congressional Research Service, warned Sensenbrenner's committee last month that NASA's attempts to produce last-minute contingency plans to replace the Russian service module are "only one example of the technical confusion surrounding the space station today". **Tony Reichhardt**

US scientists rally around call for seven per cent

[WASHINGTON] Professional societies representing more than a million US scientists, engineers and mathematicians have joined forces to call for a budget increase "in the range of seven per cent" for research at all major science-funding agencies in the federal government.

The statement, addressed to President Bill Clinton and the Congress, and due to be released at a press conference on Tuesday (4 March) of this week, is signed by 23 societies, including the main organizations representing physicists, chemists, earth scientists, mathematicians, and electrical engineers.

But biomedical researchers, who are campaigning for a larger increase in funding for the National Institutes of Health (NIH), did not join in the exercise, fracturing the otherwise united front which the societies are seeking to present.

Nonetheless, Allan Bromley, president of the American Physical Society (APS) and dean of engineering at Yale University, said that the community had achieved "unprecedented unity" in agreeing on the joint statement. Bromley says he is "increasingly optimistic" that message will be heard: both houses of Congress, he says, are starting to view investment in science more favourably.

The joint statement says that increases of around seven per cent would "strike a fair balance between current fiscal pressures and the need to invest in activities that enable

long-term growth" and that they would "partially restore the inflationary losses that most agencies have suffered during the last few years".

Arthur Jaffe, president of the American Mathematical Society (AMS), said that the budget for the 1998 financial year proposed by President Bill Clinton (see *Nature* 385, 565; 1997) does not keep up with inflation, and is therefore "inconsistent with president Clinton's goals" as stated in the president's State of the Union address. Clinton's budget, which will be considered by the Congress between now and October, calls for increased research spending of around two per cent.

The joint statement springs from an effort which was initiated late last year by the American Chemical Society (ACS), the Federation of American Societies for Experimental Biology (FASEB), and the APS to better co-ordinate their lobbying activities (see *Nature* 384, 393; 1996).

But according to sources at several of the societies, FASEB, which represents biomedical researchers, had wanted to restrict the joint initiative to support for the National Science Foundation. The organization is said to have been concerned that a call for increases across all agencies was unrealistic, and that the call for seven per cent would undermine its own campaign for a nine per cent increase at the NIH. **Colin Macilwain**

India proposes first increase for six years

[NEW DELHI] In India's first 'science-friendly' budget for six years, the government has proposed increasing spending on science by 11 per cent next year, and has launched a scheme to encourage outstanding Indian scientists to carry out world-class research.

The budget for 1997-98, placed before parliament last week, also gives a boost to research in industry, while providing money to modernize the 40 laboratories of the Council of Scientific and Industrial Research (CSIR).

India's scientists will have about US\$200 million more to spend on research than this year. The research community has welcomed the budget, which provides \$1,760 million for science-related activities, as well as \$192 million for nuclear-power schemes.

In addition, the government is proposing to set aside \$14.2 million "to assist outstanding scientists under the age of 45 to attain and sustain world class science". The move, intended to stem India's brain drain, was prompted by concern about "declining

interest in sciences in school and colleges".

As well as an across-the-board increase in funding for all scientific departments, the government has offered to match any private-sector funding raised by the national laboratories.

In future, publicly funded research institutions will be able to invest their knowledge as equity in private-sector companies. Scientists have also welcomed a decision to reduce customs duty and waive excise duty on scientific equipment.

As in previous years, the largest parts of the science budget will go on space, atomic energy and defence research. The space department has been promised funds to start work on a direct-to-home broadcasting satellite system.

But oceanographic research has won a 75 per cent boost. In the area of health, funds for malaria control have gone up by 30 per cent while support for AIDS prevention will drop from \$40 million last year to \$28 million this year. **K. S. Jayaraman**

Calls for cloning ban sell science short

Declan Butler and Meredith Wadman

Talk of restrictions on human cloning has failed to take account of the possible medical benefits and to consider public attitudes in depth.

Scientists on both sides of the Atlantic last week warned political leaders against excessive haste in legislating on human cloning, saying that overrestrictive measures could impede both important research and progress in animal biotechnology and medicine.

The warning follows the wave of concern expressed by political and religious leaders worldwide about the first successful cloning from an adult cell of a mammal. The clone, Dolly the lamb, was produced by Ian Wilmut and colleagues at the Roslin Institute and PPL Therapeutics near Edinburgh (see *Nature* 385, 810; 1997).

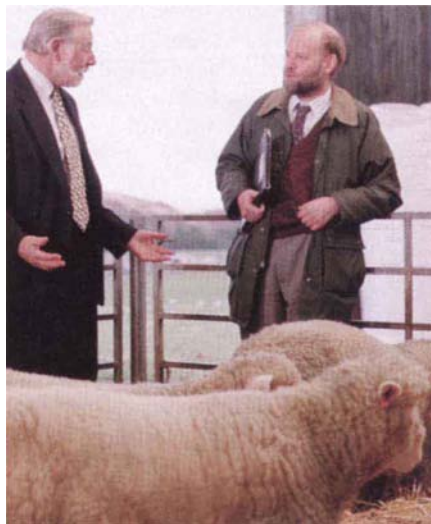
President Bill Clinton instructed the National Bioethics Advisory Commission to report to him on the issue within 90 days. Jacques Chirac, the French president, Jacques Santer, president of the European Commission, and Federico Mayor, director general of Unesco, also instructed their respective bioethics advisory committees to provide advice on what to do.

Calls for an outright ban on human cloning came from quarters ranging from the Vatican — arguing that human beings have a right to be “born in a human way, and not in a laboratory” — to the US Biotechnology Industry Organization, which represents more than 700 companies. Members of the European and British parliaments also called for a moratorium on all research on animal and human cloning. Philippe Vasseur, France’s farm minister, warned of the spectre of “six-legged chickens”.

Cloning ban could stifle research

But Harold Varmus, director of the US National Institutes of Health (NIH), last week warned a congressional subcommittee against legislating “precipitously” on cloning techniques, as this might result in overrestrictive legislation that could impede new avenues of research (see box, opposite page).

The strength and tone of much of the political reaction to the cloning breakthrough has been met by researchers with a mixture of frequent bemusement and occasional dismay. While acknowledging the need for ethical debate, many scientists argue that politicians and the media are focusing excessively on futuristic — often hypothetical — applications, to the detriment of serious debate about the more



Great debate: Ian Wilmut (right) talks to Terry Sink, a senior animal technician at Roslin.

immediate and realistic possibilities.

Guy Pailloton, director general of the French agricultural research organization, INRA — who says he is “resolutely opposed” to human cloning — appears to speak for many when he argues that “stupid talk” about human cloning has been “whipping up panic”, obscuring the real scientific issues.

Coverage of the issue by the mass media has been “depressing” says Steve Jones, professor of genetics at University College London and a prominent popularizer of science. He says his “heart sank” while watching a British television news broadcast consisting of “Dolly the lamb, Adolf Hitler and Frankenstein”. Politicians and the media have shown their “inability to differentiate between the plausible and the impossible”.

Human cloning is “certainly not for tomorrow”, and may never be a practical proposition, claims Derek Burke, a microbiologist and former vice-chancellor of the University of East Anglia. He points out that Dolly was the only lamb to result from 277 fusions involving adult cells, that only eight lambs survived from a total of 834 fusions, and that no information is yet available on Dolly’s longevity or fecundity.

Burke points out that no ethics body would approve human research carrying such risks of deformation or early death. At the same time, these hurdles are unlikely to be overcome without human experiments.

This Catch-22 could effectively bar development of human cloning, he says. At the very least, it suggests that publicly accountable institutions in the West are unlikely to embark on human cloning in the near future (although medical ethics may not deter wealthy or powerful individuals from doing so out of the public eye).

The cloning debate therefore needs to put such concerns in perspective with the more immediate and realistic repercussions of the breakthrough, says one scientist. Cloning techniques should allow rapid progress in basic research, such as in the understanding of what switches genes on and off during development, and in the production of animal models for studying human diseases. Animal cloning could bring advances in agriculture, facilitate the production of pharmaceuticals in animal milk and help to conserve endangered species.

Scope for medical advances

Wilmut’s group and PPL Therapeutics are themselves trying to produce in sheep’s milk large quantities of alpha-1-antitrypsin, a compound that inhibits elastase, an enzyme that damages the lungs and pancreas in cystic fibrosis.

A precipitous ban on all research on human cloning could prevent the development of life-saving medical treatments, according to several scientists. Varmus last week told the labor, health and human services and education subcommittee of the House of Representatives appropriations committee that funds NIH that cloning techniques could be used to generate skin grafts for burn victims and bone marrow for patients undergoing cancer chemotherapy.

But if many scientists are concerned about the threat of politicians imposing overrestrictive legislation, they are also aware that rushing into research on human cloning could prove counterproductive. It could spark a public backlash resulting perhaps in even stricter legislation that would stifle cloning research, and perhaps genetics research in general.

“People feel strongly that we are moving

GREAT MOMENTS IN GENETICS: THE CLONING OF KING CANUTE.



into areas where they do not want to go," says Burke, adding that many view cloning as "an invasion of personality" and "a Pandora's box with unpredictable consequences". He believes scientists need to first let the public become accustomed to cloning of animals. "We should not move on [research into] human cloning; society is not ready for it."

Jones says: "The public is not frightened of progress but of rapid progress." He points out that the once controversial techniques of test-tube babies and organ transplants are now widely accepted. The job of ethics committees, he says, is to act as a "brake", slowing the application of technology to a speed acceptable to the public.

Several observers predict that human cloning will shift from being "totally unacceptable" to being permitted, say, for tissue culture or even treating some forms of infertility. Experience suggests that public opinion may support a mother who wishes to clone her dead child, says one scientist. "It is possible that after a long debate society might accept cloning under controlled conditions," says Burke.

Many observers predict that the passionate debate of the past few days is likely to subside relatively rapidly. The cacophony of calls for a ban on all research on human cloning looks set to be followed by a realization that, in a democracy, legislators usually need good reasons for restricting scientific progress.

Does cloning mean new ethics?

Beyond concern that science is moving so quickly that society cannot keep up, the arguments of proponents of a ban on animal and/or human cloning have generally not gone much deeper than vague assertions that these pose "new ethical problems" or are "contrary to nature". The various ethics committees now convened on cloning should bring some welcome clarification.

The 'contrary to nature' argument fails to take into account "the reality that man has long manipulated animals to his own ends", says one observer. Jones says: "I felt like asking Clinton; do you eat lamb chops?"

The widely held philosophical argument that human cloning attacks the fundamental principle of the 'human dignity' of individuals is also challenged by some. This claim, they argue, amounts to an apology for genetic determinism, in that it implicitly ignores the influence of 'nurture' on personality. Identical twins are given two souls by the church, and two votes by the state — not one.

Indeed, scientists have strongly emphasized that human clones would not be strictly 'identical', as they would be born a generation apart and experience different environmental influences. A clone of Hitler would not necessarily become a dictator; and a clone of one's father would not actually be one's father, as his relationship would not be a parental one.

According to Burke, the frequency with which such reservations have been expressed in the debate about cloning itself marks a significant shift away from the recent trend towards genetic determinism. "It has shifted the balance back towards the nurture end of the nature-versus-nurture debate."

Others argue optimistically that society's collective ideological apparatus is sufficient to prevent abuses of cloning. One widely aired concern is the prospect that human clones might be created as a reserve of spare parts, but nature produces identical human twins, and no one has yet converted one into an organ bank for the other. Humanity would also be unlikely to abandon the benefits of sexual reproduction — genetic diversity and hybrid vigour — in favour of cloning.

What most people seem to agree is that the main risk of human cloning is that it would allow third parties to impose biological

predetermination. They argue that the lottery of heredity affords the best protection against this threat.

But at least one scientist, Richard Dawkins, professor of public understanding of science at the University of Oxford, confesses to a desire to be cloned. "I think it would be mind-bogglingly fascinating to watch a younger edition of myself growing up in the twenty-first century instead of the 1940s."

Indeed, a wider outcome of the debate about cloning is that it has prompted some scientists to warn against what they claim is an increasingly unjustified stigmatization of the potential dangers of genetic research. "I find it astonishing that people complain bitterly about the perceived threat of genetics to human dignity when they see grotesque insults to human dignity all around them which could be changed tomorrow and they do nothing about", says one scientist. □

Putting the lid on Pandora's box of genetics

[PARIS & WASHINGTON] The United States and the European Union have both commissioned immediate reviews of the ethics of cloning research.

But it is far from clear what regulations — if any — will be imposed on animal cloning, or indeed on research into human cloning. Nor are there any obvious ways of drawing a strict borderline between research in the field that is and is not ethically acceptable.

That the world's legislatures were so unprepared for the advent of cloning can be explained in part by the widespread belief that the ability to clone mammals from adult tissue was a long way off.

"Gobsmacked" is how several geneticists describe their reaction to the news.

Both Francis Collins, director of the US National Human Genome Research Institute, and Bruce Alberts, president of the US National Academy of Sciences, concede that they were "totally caught off guard" by last week's news. "It's hard to get [busy experts] to spend a lot of time on hypothetical scenarios that seem like they're almost certainly not going to come true," says Alberts.

In the United States,

research into cloning could currently proceed in the unregulated private sector of reproductive medicine. One option that the National Bioethics Advisory Commission will consider is whether curbs on this sector are needed. Public research agencies funded by the Department of Health and Human Services — including, most prominently, the National Institutes of Health — could not carry out cloning research as it would fall under an existing ban on federal funding for human embryo research. But technically, if not politically, research could currently be funded by other government departments.

A New York state senator, John Marchi (Republican, Staten Island), has introduced a bill in that state's legislature that would make human cloning a crime carrying a three- to seven-year prison sentence. The bill would also give the New York State Department of Health regulatory authority over research on animal cloning. But the governor of New York, George Pataki (Republican), said it was "just too soon" for politicians to jump into the fray.

The European Commission's advisory group on biotechnology will carry

out a broad review of genetics research and see whether regulations need tightening, according to its chairwoman, Noëlle Lenoir. She is a member of the French Constitutional Council and chairwoman of Unesco's International Bioethics Commission. Work on human cloning is already forbidden in research funded by the European Union, she points out.

Jürgen Rüttgers, Germany's science minister, has sought to reassure the German population, pointing out that the cloning of people is strictly forbidden under the country's 1990 embryo protection act. Meanwhile, Ernst Benda, a former president of the German Constitutional Court has criticized Unesco's draft convention on bioethics for failing to explicitly forbid human cloning.

The House of Commons Select Committee on Science and Technology has agreed to meet to discuss whether existing UK legislation contains loopholes that might allow human cloning (see *Nature* 385, 767; 1997). The French national bioethics committee is to study the Country's bioethics laws which are considered to ban cloning, but do not do so explicitly. **D.B. & M.W.**

Hughes widens funding for research projects to South America

[WASHINGTON] The Howard Hughes Medical Institute, the largest source of private funds for biomedical research in the United States, announced last week that it is expanding its reach to include projects in four further countries in South America.

The institute has awarded \$15 million in grants to 47 scientists in Argentina, Brazil, Canada, Chile, Mexico and Venezuela. Research being funded includes studies of the genetic origins of cancer, the molecular basis of parasitic diseases and mechanisms of cell growth control.

The awards mark an expansion of the institute's international research programme, which began in 1991 when \$11 million was awarded to scientists in Canada and Mexico. In 1995, it supported researchers in Belarus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Russia, the Slovak Republic and Ukraine.

Galileo given longer look at Jupiter's moons

[WASHINGTON] Galileo spacecraft managers at the Jet Propulsion Laboratory (JPL) in Pasadena, California, have been given

permission to extend their exploration of Jupiter by another two years. Originally the mission was to end in November. But last summer JPL proposed a low-cost extension to the National Aeronautics and Space Administration (see *Nature*, 382, 745; 1996). For an additional \$30 million, Galileo will fly past Jupiter's moon Europa eight more times, and Calisto four more times, ending with a close fly-by of the volcanic moon Io. The money will come from savings resulting from improved efficiency in spacecraft operations, according to Galileo managers.

Danish drug company buys Indian products

[LONDON] In what is said to be the first decision by an Indian company to license its discoveries to an international pharmaceutical company in return for advance payments and royalties, the Danish pharmaceutical company Novo Nordisk has bought the rights to develop and market pharmaceutical products manufactured by Dr Reddy's Research Foundation, a group of companies based in India.

Novo Nordisk is known for its diabetes products. Bruce Carter, an executive vice-president of the company, said the agreement with the foundation, which has an annual turnover of US\$100 million, will enable his company to gain access to

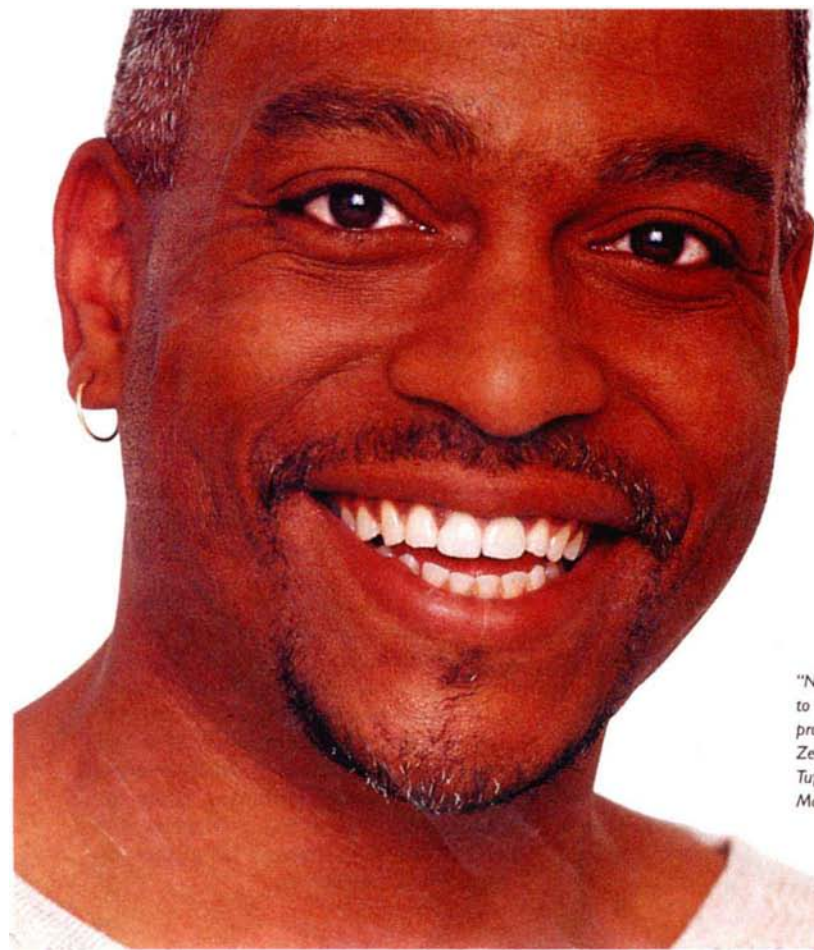
compounds that may "prove useful in the treatment of diabetes and its complications". The agreement covers compounds thought to have potential for treating diabetes, obesity, dyslipidemia and related complications.

German industry spends more on research

[MUNICH] After several lean years, German industry has started to increase its spending on research and development (R&D). Statistics published last week show an estimated rise in spending of 2 per cent in 1995 and 1.8 per cent in 1996. The trend seems likely to continue; a quarter of manufacturers predict higher spending on R&D this year compared to 1996. But the number of staff employed in industrial R&D continues its downward path. Since 1991 the number of research workers has fallen by 15 per cent to 274,000.

FDA plans guidance on human cell procedures

[WASHINGTON] The US Food and Drug Administration issued a plan last week for regulating human cells and tissues used in procedures ranging from tissue transplants to gene therapy that is intended to replace the current patchwork system of regulation.



Malcolm is imp
by the new Pre
Protease be
it's a GST
fusion prot

"Not being a protein chemist, I just want to clone the gene, express it, isolate the protein and move on," says Malcolm Zellars, who's working on his post-doc at Tufts University Medical School in Boston, Massachusetts, USA.

The proposed framework adopts a graded approach to regulation depending on the degree of risk of contamination or damage posed by cells or tissues. Gene therapies, for example, would fall under the strictest category of scrutiny, requiring pre-market approval and controlled clinical trials just as new drugs and medical devices do now. At the other end of the scale, tissues removed and transplanted to the same patient in a single procedure would not be regulated.

Basic research is priority for German universities

[MUNICH] Klaus Landfried, who was last week elected president of the German University Rectors Conference, says that defending basic research in German universities will be one of his main priorities when his three-year period of office begins in August. With federal and regional governments preferring to concentrate their shrinking research funds in a few areas of applied research, Landfried warns that small university groups, whose research does not fit into the chosen categories, could be the biggest losers. Landfried, a professor of political science at the University of Kaiserslautern, near Mannheim, promises to fight for more money from the federal government to repair university buildings, particularly in the neglected east of Germany. Many research

buildings are in such a dire state that scientists cannot comply with safety regulations.

Oscar nominations for US science films

[WASHINGTON] Special Effects and Cosmic Voyage, two educational films made with the support of the US National Science Foundation, have been nominated for this year's Oscar in the documentary, short subject category by the Academy of Motion Picture Arts & Sciences. The films, which were made in the large-screen, IMAX format, deal respectively with the use of science and mathematics in movie-making, and with the size and scale of the Universe.

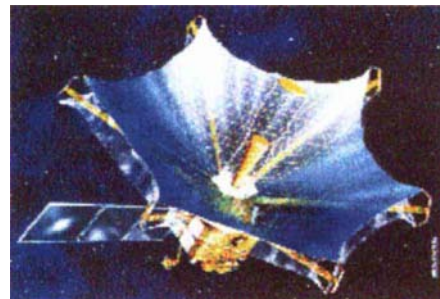
French law ruling queries Superphénix decision

[PARIS] Environmental groups last week won a legal challenge over the French government's decision in 1994 to restart Superphénix — the world's largest prototype fast-breeder reactor — not to produce power station but to study the incineration of plutonium (see *Nature* 370, 401; 1994).

The Council of State, the supreme court, ruled that the public inquiry that preceded the 1994 decision was based on the use of the reactor — at Crey-Malville near Lyon — for generating electricity and could not be

considered as a proper evaluation of the reactor for research purposes. The most likely outcome is that the government will be obliged to hold a new public inquiry.

Astronomical antenna unfolds in space



[TOKYO] The world's first programme in space-based radioastronomy passed a critical step last week with the successful opening of the 8-metre parabolic antenna (above) on the Japanese radiotelescope satellite HALCA (Highly Advanced Laboratory for Communications and Astronomy). The satellite was launched last month to carry out observations in Very Long Baseline Interferometry (VLBI) in conjunction with ground-based telescopes (see *Nature* 385, 663; 1997). Known also as VSOP, standing for VLBI Space Observatory Programme, the satellite will begin test observations in April.

ressed
Scission
cause
ein

Are you working with GST fusion proteins? A new one-step solution offers you complete simplicity in purification. Introducing pGEX-6P vectors and a recombinant GST fusion protein called PreScission™ Protease.

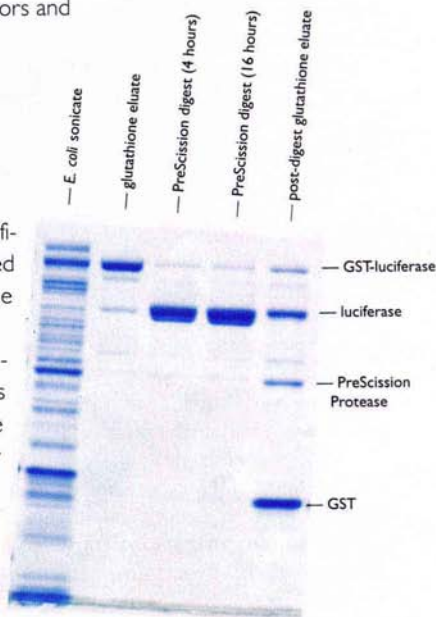
One-step separation of the protease and GST affinity tag from your protein of interest

Following the site-specific cleavage of GST fusion proteins, purification of the protein free of PreScission Protease, and the liberated GST affinity handle, are accomplished in one step while using the same affinity medium, Glutathione Sepharose®.

The revolutionary PreScission Protease is similar to other commercial site-specific proteases in that its recognition sequence is rarely present in most native proteins. What's more, the protease performs at a 5°C temperature optimum to preserve the integrity of your protein. This quick and easy, single-step solution is only available from one company, Pharmacia Biotech.

Are you doing gene expression?

Give us a call: 1 (800) 526-3593 in the USA; 03 3492 6949 in Japan; +46 18 16 50 11 in Europe and the rest of the world. Or visit us on the Internet: <http://www.biotech.pharmacia.se>.



**Pharmacia
Biotech**

Uppsala, Sweden. (And the rest of the world)

Flawed reasoning about oil and gas

Sir — The review by Peter Kassler¹, the book he reviewed² and other similar works of recent years perpetuate an incorrect and insidious myth about the quantity and accessibility of petroleum and natural gas resources. This myth essentially claims that the world will run out of oil and gas at some future time because the resource is limited. Whereas any specific resource defined in a limited manner is indeed limited, concerns about imminent long-term shortages of oil and gas are flawed for several good reasons.

The amount of hydrocarbon available to mankind is not limited to that from conventional oilfields. Less than 5 per cent of the hydrocarbon in sediments worldwide, exclusive of coaly matter, is in the form of conventional oil and gas. Vast resources of heavy (viscous) oil exist, not only in Canada (250 billion cubic metres) and in Venezuela (300 billion cubic metres) but also in similar deposits in the Middle East, China, Russia and elsewhere.

The known amount of heavy oil is substantially larger than the amount of conventional oil, perhaps by a factor of two or three. These are seldom, if ever, included in the assessments of oil supplies.

Technological advances using horizontal wells, gravity drainage with thermal stimulation and other new production approaches³ are rapidly lowering the costs of these resources. In Canada, operating costs have dropped from C\$70–\$80 per cubic metre in the late 1980s to C\$25–\$40 in 1996 for these technologies. Also, the exploration costs associated with these resources are essentially nil, whereas the exploration costs for conventional oil continue to rise inexorably (C\$40–\$50 per cubic metre in the Western Canadian Sedimentary Basin).

The implications for petroleum supplies are substantial. Consider the consequences if only 20 per cent of the Canadian heavy oil

deposits are economically recoverable (an estimate I consider to be slightly pessimistic). This amount is sufficient to meet current US and Canadian consumption combined for a period of more than a hundred years. And Canadian “political and military disruptions” are modest, compared to recent Middle East events.

Despite being a political football, oil remains essentially a commodity. As conventional oil deposits are gradually depleted, prices will rise, and this will make non-conventional resources economically accessible. One may ask at what cost. In Europe, 60–80 per cent of the commercial cost of a litre of fuel is government-imposed taxes. A simple calculation shows that a commodity price of US\$600 per cubic metre could be sustained without an increase in price at the pumps in Europe, if the governments involved are willing to forgo this lucrative source of revenue in the distant future. US consumers who benefit from low fuel taxation will have a different opinion on such a commodity price.

What is the consequence of rising prices for a commodity-priced barrel of oil? First, resources that have been considered marginal (oil sands, smaller offshore fields) become economically exploitable. Second, fields that had been abandoned after producing only 20–40 per cent of the oil in place with conventional technology will be redeveloped using gravity drainage methods to access a large proportion of the remaining oil. Third, if prices are sufficiently high, with some confidence of price stability, other vast resources become available for development. These include, for example, oil shales and coal.

It is estimated that 90 per cent of the non-coal hydrocarbons in sedimentary basins are found in shales. The Green River oil shales (actually kerogenous shales) of

Colorado and Wyoming are but one example of such deposits, and the estimated hydrocarbon volume in this deposit is similar to that in the Canadian oil sands.

It is also possible to make synthetic crude oil through the hydrogenation of coal: the Sasol process in South Africa is the prime international example of such technology. Vast quantities of coal and low-quality lignite exist throughout the world for this purpose.

Finally, given a sufficient commodity price, one may even extract carbon (from atmospheric carbon dioxide or limestone) and hydrogen (from solar-powered electrolysis of water) and assemble synthetic gasolines from the basic elements. The only realistic limitation to this process is price.

The view, therefore, that hydrocarbon resources are inherently fixed or limited is false. Government policies are based on the politics of conventional oil, and these policies will be revamped as technologically intensive non-conventional sources of oil become, unavoidably, an increasing part of the commodity supply.

Limitations on oil use are therefore more logically related to environmental issues such as global warming and urban pollution; resource limits do not, for practical purposes, exist. Oil shortages are actually short-term shortfalls in cheap conventional crude oil supplies, and have little to do with the actual long-term hydrocarbon supplies.

Maurice B. Dusseault

*Porous Media Research Institute,
University of Waterloo, Waterloo,
Ontario N2L 3G1, Canada
e-mail: mauriced@sciborg.uwaterloo.ca*

1. Kassler, P. *Nature* **384**, 528 (1996).

2. Collett, M. & Simeoni, C. *Perspectives of Oil and Gas: The Road to Interdependence* (Kluwer, 1996).

3. Dusseault, M. B., Geilikman, M. B. & Roggensack, W. D. in *Proc. Soc. Petroleum Engineers Heavy Oil Sym.*, Calgary, Alberta (Publ. no. 30250, Soc. Petrol. Eng., 1995).

Roslin unfunded

Sir — The work of Ian Wilmut *et al.* (*Nature* **385**, 810–813; 1997) from this institute on nuclear transfer raises a number of general issues about the nature of research and its exploitation and funding, particularly in the United Kingdom. Biological research is going on in hundreds of laboratories, proceeding in the main by incremental steps. Not even scientific experts in a field know who will take the next step. This was exactly the case for our nuclear transfer research. Several laboratories in different countries were working on the same problem. My colleagues had one clever idea which put them ahead of their competitors — if we had stopped, someone

else would have done it sooner.

Scientific progress cannot be easily regulated, but its application can. We see no reason for applying this technology to humans. But, of course, we see tremendous opportunities for application to animal biotechnology, whether for the production of valuable human proteins in animals for medical purposes or to maintain the competitiveness of our animal breeding industry.

Most of the funding for our nuclear transfer project came from the Ministry of Agriculture, Fisheries and Food, which has decided to withdraw further support. This decision raises the issue of the funding of animal biotechnology and similar strategic

research in the United Kingdom. Although the UK Technology Foresight Plan provides clear objectives for strategic research, there is no adequate framework to implement them. We need a 5–10 year implementation plan for animal biotechnology to which relevant government funding agencies, as well as industry, can contribute. This might avoid future funding problems for high priority projects that fall between, or overlap, the policy objectives of different funding agencies.

Grahame Bulfield

*(Director and Chief Executive)
Roslin Institute,
Roslin, Midlothian EH25 9PS, UK
e-mail: grahame.bulfield@bbsrc.ac.uk*

Footloose managers run away from research

Sir — Scientific research within broadly based industrial companies has nearly vanished in recent years. Among the major factors contributing to this decline is the relatively short tenure of high-level industrial managers. Although the effect of tenure has often been mentioned, few quantitative data have appeared. Here I give some well-founded data for one company, General Electric (GE), and limited, less perfect data for IBM and AT&T.

GE annual reports to stockholders list the high-level management (vice-presidents and above) and their positions — typically about one hundred. A chronological sequence of such reports supplies the overall attrition rate of people who are shifted, promoted, hired elsewhere, retired or fired. Half-lives τ can be found from the starting number N_0 and the ending number N_t by assuming that the fractional attrition over each interval of time t is constant ($\tau = t \log_{10} 2 / (\log_{10} e) \ln(N_0/N_t)$).

Over the years 1981 to 1995, half-lives ranged from 1.5 to 3.4 years, with an average of 2.3 years and a median of 2.1 years. From 1975 to 1977, the two annual values were 1.8 and 2.8 years, based on listings in Standard and Poor's annual *Register of Corporations, Directors, and Executives*. These data are in general less reliable because listings are not necessarily updated annually. Intervals with zero attrition are therefore rejected. Standard and Poor's listings for AT&T (1988, 1989, 1993, 1994 and 1995) and IBM (1989, 1990 and 1991) gave various half-lives between 1 and 3 years — similar to those at GE.

It is unlikely that most managers whose promotion or lack of it will be decided within two years of beginning a job will support research work that can provide useful products or processes only over a longer interval. This mass of high-level management with brief tenure forms a cadre that can be expected to provide continuous pressure toward short-term technological endeavours that will tend to emphasize quick payoffs by short-term engineering and development activities and to discourage expenditures on research aimed at long-range needs and opportunities.

It may appear paradoxical that long-term science ever existed in companies where management tenure is so brief. The answer is that management is hierarchical rather than democratic, and a look at the data shows that the highest management is far more stable than are the next lower layers. In the present case, the GE annual reports show that the 1981 to 1995 half-life

of management was 5.1 years for the roughly 30 members at the corporate headquarters, and one person, John F. Welch, has been president throughout (and four previous presidents averaged 9 years in office before retiring).

Given that the chief executive of a corporation can often expect long-term tenure, his or her decisions are likely to reflect needs well into the future. If that person concludes that research will further the future of the company, that activity will flourish even in the face of opposition by the bulk of the less securely placed management. If the chief executive tilts in the opposite direction, there will be ample internal support.

Robert L. Fleischer

*Department of Earth
and Environmental Sciences,
Rensselaer Polytechnic Institute,
Troy, New York 12180-3590, USA*

One for the fairies?

Sir — There is something rather endearing about your second, brief leading article "Avoid financial 'correctness'" (*Nature* 385, 469; 1997); perhaps its charm and innocence would have been enhanced by an Arthur Rackham illustration in which fairies danced around toadstools?

Is there indeed any "point" in authors of scientific papers declaring their business interests in the topic under discussion, when these authors are upright, thoroughgoing, bona fide scientists, and not just ordinary human beings? Apparently not.

There are, of course, a few case histories on the matter, including those research



specialists who were unable to find a link between smoking on the one hand and lung cancer and heart disease on the other, when those same specialists were paid by disinterested and unbiased tobacco companies; a number of other cases involving drug companies footing the scientific bills for careful and meticulous analysis of their drugs also come to mind.

In many such cases, as I recall, the source of funds, grants and salaries for the scientists to conduct their research was not unduly emphasized, even by the scientists themselves. Cynics might think that a financial interest could conceivably play some role.

It is therefore reassuring to those of us who have spent our lives locked in an attic that *Nature* will have no truck with such base suspicions and "will persist in its stubborn belief that research as we publish it is indeed research, not business", funny, monkey or anybody's.

Remember, every time someone says "Declare business interests", a fairy dies.

Ralph Estling

*The Old Parsonage,
Dowlish Wake,
Ilminster,
Somerset TA19 0NY, UK*

No Greenpeace 'hijack'

Sir — Your leading article "Risk and the inadequacy of science" (*Nature* 385, 1; 1997) claims that, in the case of Brent Spar, Greenpeace "hijacked" the issue of risk, "making the engineering assessments of risk undertaken by Shell effectively redundant as far as the public were concerned".

This is to misstate our case. In the case of Brent Spar, our main objectives were to prevent a precedent for dumping oil/gas installations at sea and to ensure that industry took responsibility for dealing with the waste it produced.

Although risk was relevant, a conventional risk assessment was an inappropriate tool to resolve the issue, as the main questions to be contested lay outside it. One relevant factor was that of need. We argued that there was no technical engineering need to dump at sea, and this has now been borne out by the fact that Shell has shortlisted 11 options for the future of Brent Spar, none of which involves deep-sea disposal.

The public were right to sense that the framework of decision-making was wrong and redundant. That is not to reject science or engineering assessments: it merely argues for their appropriate use.

Chris Rose

*Greenpeace UK,
Canonbury Villas,
London N1 2PN, UK*

Humour should be taken seriously

Sir — P.-L. Chau and W.Y. Chau ignore some important points in their defence of Confucianism (*Nature* 385, 110; 1997).

First, Confucianism as actually practised, emphasizing deference towards seniors and elders, has resulted in authoritarianism (as well as such positive virtues as refinement of the person). Second, Confucian-influenced cultures such as China, Korea, Japan and Vietnam were not deeply permeated by Christianity and its concomitant emphasis on the importance of the individual. This distinctive Christian individualism underlies the Weber/Tawney thesis that Christianity gave rise to capitalism. It has also greatly encouraged originality in the West, because the individual who thinks independently of the group or tradition in which he or she lives is usually respected and valued.

Japan has suffered from the same absence of scientific creativity as has Korea. But in China, Taoism, with its emphasis on spontaneity and paradoxes, probably encouraged more original scientific thought, particularly in the fields of electromagnetism and the invention of the compass. Nevertheless, China did not undergo a scientific revolution, as did the West in the sixteenth and seventeenth centuries, probably because it lacked Christian-influenced individualism.

Confucianism does encourage self-discipline and is excellent for arduous training techniques. Korea and Japan excel, for example, in the efficient production of players of musical instruments. It could, however, be argued that there are now too many Korean child prodigy violinists in the West. An original and beautiful musical composition from Korea or Japan would surely now be more appreciated.

In a leading article on Confucianism (*Nature* 384, 197; 1996), you ask "So how to boost scientific creativity?". One important way to do this, hitherto ignored, is by encouraging humour in scientific communities which often think convergently. Humour develops divergent thinking, an important aspect of creativity. Wit can often shift a train of thought to an entirely new mode of perception. Above all, humour usually involves the momentary fusion of two habitually incompatible frames of reference. The sudden comprehension of abrupt incongruity generally results in laughter and the release of tension. It frees the mind from the restraints of logic and rational decision. The resulting relaxed playful atmosphere often encourages further divergent and creative thought.

The precise relationship between

creativity and humour has not been given the attention it deserves. In his MSc thesis (University of Surrey, 1995), Vaughan Goddard demonstrated that subjects who watched a videotape of Stephen Hawking's *A Brief History of Time* before taking a creativity test scored lower than a group of people (with similar backgrounds) who had just watched a humorous videotape (Jack Dee).

Humour should therefore surely be regarded very seriously by all scientists, even including those from non-Confucian cultures.

Audrey Wells

*Department of History,
Royal Holloway,
University of London,
Egham, Surrey TW20 0EX, UK*

The 's' in Unesco

Sir — Declan Butler's News article about Unesco (*Nature* 385, 286; 1997) omits any mention of the major issues that confront any intergovernmental approach to science. As a former assistant director general (science) at Unesco, I should like to make the following comments.

The main issue probably relates to the inextricable relations between science and technology, that is, between free research for new knowledge and jealously patented intellectual property. This leads to diminishing support from governments for basic science. This also explains why developing countries have constantly pressed for a better share of 'science and technology', through, among other things, the UN conference on the subject held in Vienna in 1979, without ever obtaining anything but nice words.

Science itself lies in the minds of scientists and is essentially concentrated in the industrialized world. It has its own traditional channels of communication, particularly through the multiple unions and associations gravitating around the International Council of Scientific Unions. Specialists in brain research, or indeed in physics or chemistry, do not necessarily find much value in cooperating through Unesco, and most of them do not feel much interest in promoting their disciplines in the developing world. But the developing world needs some means of contact and participation in scientific matters if it is actually to develop. Weakened as it may be because of the absence of the United States and the United Kingdom, and because of lack of support from international funding agencies, Unesco is still the only multilateral organization devoted to the promotion of science in the developing world. But obviously it cannot do so without support from the industrialized countries.

These countries are quite happy to use Unesco when it suits their own immediate interest as, for instance, when a world forum is needed in bioethics, or when the nature of the research work to be conducted implies the goodwill and participation of developing countries. This is clearly true when the geographical dimension of the problems is paramount. As you rightly point out, the international scientific programmes of Unesco in hydrology, geology and oceanography are "well respected". To these could be added programmes in microbial resources and in biological diversity, and particularly the Man and the Biosphere Programme, with its world network of 'biosphere reserves'. Although they do not contribute to their overhead costs, the United States and the United Kingdom continue to take part in and to benefit from these global research programmes. This fact alone seems to justify the value of the 's' in Unesco.

Some multilateral mechanism is more than ever needed to ensure better access to science on a worldwide basis and to deliver cooperatively the necessary knowledge on global interdisciplinary issues which govern our future on this Earth. Unesco has received this mandate. All those concerned should see to it that it fulfils it well.

Michel Batisse

*7 Rue de la Cavalerie,
75015 Paris, France*

Science and emotion

Sir — Few objective scientists would disagree with the view that emotion and passion have a very real place in the world of science. Until fairly recently, however, this idea might have occasioned an outcry from many scientists and would certainly have been unlikely to appear in the editorial columns of *Nature* (385, 373; 1997).

But it is largely because of so-called 'science studies' that the myth of the passionless scientist is fast disappearing. Although it may be true that the passions of science "... may often have little impact on what eventually becomes accepted as scientific truth", the process of reaching that goal, as I have attempted to demonstrate in a recent book (*Envy as a Retarding Force in Science*, Avebury, Aldershot, 1996), may well be slowed and retarded by the human emotions that are often at play in science. Indeed, envy may be one of the important human emotions responsible for the inertia that often delays the acceptance of profoundly important scientific truths.

Mark A. Gillman

*South African Brain Research Institute,
6 Campbell Street, Waverley,
2090 Johannesburg, South Africa
e-mail: mag@iafrica.com*

Contending for the dead

Should archaeological remains be returned to the cultures from which they were removed? We should consider the needs of researchers and wider society as well as acknowledging the sensibilities of indigenous communities.

**D. Gareth Jones
and Robyn J. Harris**

Recent, widely publicized arguments about the repatriation of human skeletal remains from museum and other collections highlight the opposing forces of Western scientific values and a global cultural renaissance among indigenous populations. Some see these two perspectives as implacable opponents, and believe that the destruction of ancient evidence of indigenous communities is no better than European cultural imperialism. On the other hand, physical archaeology is increasingly seen as a discipline under threat, with scientific freedom compromised by repatriation.

Archaeologists wishing to work on material today are not responsible for atrocities committed more than 100 years ago — they cannot be implicated in the unethical activities of former times, and even if they benefit from them, there is no moral connection between the two events. What is important, in our view, is the rationale for the proposed contemporary scientific work, the quality of this work, and its potential value to the wider human community.

Desire for restitution

It is now almost commonplace to hear about the return of skeletal remains from anatomy departments, medical schools and museums to indigenous communities for reburial. High-profile cases have involved the return of Australian Aboriginal skeletal material by European and Australian museums and universities. One particularly contentious case involved the return of remains, estimated to have been between 9,000 and 15,000 years old, by the Museum of Victoria, Australia. In the United States, remains of Native American people have been returned for reburial by several major universities and by the Smithsonian Institute, which had a collection of 18,000 mostly prehistoric bones.

One bitterly divisive case of repatriation involved the return, in 1991, by the Smithsonian Institute of 756 sets of skeletal remains, comprising an estimated 1,000 individuals and a large collection of associated funerary objects to the people of Larsen Bay (Kodiak Island) on the southwest coast of Alaska. The bones dated back 2,000 years and had been collected in the 1930s from around 800 unmarked graves. In Israel, after lobbying by Orthodox Jews, the Government has issued a new interpretation of the Antiquities Act, which has led to the reburial of all human remains younger than

5,000 years. Some Israeli anthropologists believe that this will be the end of the discipline in their country.

These developments reflect the tension between the legitimate scientific interest in unravelling clues to the past, and the sacred beliefs of indigenous communities. Within a scientific framework, skeletal remains are part of the heritage of humanity in general, and should not be restricted to direct, let alone indirect, descendants. But many indigenous communities, whose cultures emphasize religious respect for the remains of tribal ancestors, have claimed the sole right to determine how the remains are to be treated, thereby overriding the genuine interest of others in the material. Part of the equation is an understandable desire for restitution in the face of past disrespect shown to indigenous groups by others, as well as a continuing struggle for rights and recognition. Governments are often unwilling to support science when the contentious issue of reconciliation with indigenous peoples is involved.

In assessing the various claims, we believe that respect for the beliefs and feelings of indigenous communities is no different from that which should be shown to any person or group, and that everyone's dignity is bound up with studies of the past. Demon-

strating this respect requires an acknowledgement of the autonomy of indigenous people to decide what is done with skeletal material from their own culture. Similarly, scientists must take serious note of the cultural mores of indigenous people, and acknowledge the importance of skeletal remains to such groups. Fairness suggests that account be taken of past injustices.

It follows that, when skeletal remains are recent and direct links to indigenous peoples today can be established, the interests of those groups become paramount. We believe that this should be the case because it is generally conceded that there is a close association between a body and the known person. This association, based on a knowledge of individual people and their bodies, can be extended to groups of people and their bodies. There is, however, a proviso.

In many areas of science and medicine, there are instances where the value of information derived from a particular unidentifiable individual to all humanity is much greater than any stake claimed by specific interest groups. One example is genetic information, which, when it relates directly to an identifiable individual, is confidential to that person, with consent being essential for the disclosure of that information. But when genomic information cannot be traced

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Aborigine David Mowaljarli holding a box of ancestral bones returned by Bradford University, United Kingdom, in 1990. But archaeologists and prehistorians believe that such actions will paralyse research.

directly to any identifiable person or group, the relevance of the data to the human community as a whole becomes predominant, and the ethical constraints of the former situation no longer apply. A similar argument applies to epidemiological data.

Ownership

Returning to the question of ownership of skeletal remains, two opposing viewpoints emerge. Even in circumstances where no link can be established to a particular living person or group of people, the material is still of human origin; it retains moral value, because it retains human connotations. Consequently, prehistoric skeletal remains, where no identifiable links can be established with either a direct descendant or group of descendants, are of ethical concern. However, the remains should be available for reputable scientific investigation, as the findings are, in the broadest terms, applicable to all humanity. Consent has become of secondary importance, as no living being is in a position to provide consent. What is now most important is establishing the best uses of the material to assist humanity as a whole.

So far we have limited our discussion to the treatment of remains in museums or laboratories, but what about newly discovered remains? A popular assumption seems to be that these should be handed over to living descendants for immediate reburial. But there are other options, each relying on a close working relationship between indigenous groups and scientists (and on the granting of appropriate permission). It is to be hoped that recognition of the fact that other options exist will stimulate discussion about the reasoning behind reburial, particularly when information of interest to indigenous groups themselves can be gained by scientific study of the remains.

The options should also highlight the need to specify the nature of the differences, if any, between material that has lingered in museums for many years and that which has been recently unearthed. Are there any differences? Recently uncovered material from the prehistoric past is exactly the same as material uncovered many years ago. We believe that the scientists involved in uncovering the material should have the main control over its short-term fate, as its availability is entirely dependent on their efforts.

The intellectual environment in which a great deal of extant human skeletal material was obtained is far removed from that of today's environment, characterized as it was

by enthusiasm for phrenology, and by evolutionary ideas about the relative advancement of different races. The obsession with fitting people into racial categories led to a huge increase in the collection of skulls, especially those of Australian Aborigines, who were thought to be an evolutionary link between humans and apes. The desecration of graves was commonplace, and Aboriginal people were murdered in this cause. It would perhaps be possible to dismiss these events as a ghastly historical curiosity were it not for the continued existence of this material in many universities and museums. Do the uses to which this material was put 100 or so years ago have repercussions for our view of it in the present?

According to the principle of moral complicity, those who use material or data obtained unethically are themselves implicated in the unethical practices — as if they themselves were acting unethically. One option is to reject the principle, on the grounds that return and reburial of remains only serves to damage further the heritage of these people. One can then ask what is the best use of this material to contemporary society. Alternatively, accepting the moral complicity principle leads inevitably to reburial of all skeletal material because of the inherent evil of the actions that brought about the existence of some of the collections. However, this may also lead to the reburial of anthropological material even when there is no suggestion that gross atrocities were committed.

Specific recommendations

If wholesale reburial is not advocated, then what? We cannot justify a situation where thousands of skeletal remains from the recent past linger unstudied in universities and museums for years on end. If they are in storage for scientific purposes, ongoing (even if sporadic) scientific work is essential to justify continued storage. Considerable care is required before imposing arbitrary time-limits for study before reburial, but a vague expectation of possible future work on the material is not sufficient reason for keeping it. Human material can only be kept ethically if the purposes for which it is being kept have ethical justification, and storage (except when equated with burial) fails to satisfy this criterion. Neither can educational purposes be used as justification for unlimited storage, as casts of the material are sufficient for teaching.

Where an individual or group of individuals can be identified as descendants of the human material, there has to be agreement between these people and national authorities about the storage of the material in accessible sites, and the types of scientific study permitted. Accessibility is a two-way phenomenon, serving the interests of both the scientific and indigenous communities, although pre-eminence must be given to

indigenous communities as this is their material to 'give'.

Archaeologists and anthropologists do not have pre-eminence rights to human remains and objects considered sacred by living communities. The use of human material, especially when it is from the recent past, can be ethically justified only when carried out within the context of human obligations, expectations and hopes. To ignore this context does untold damage to the cause of investigating any human material.

Where material is to be held in 'safe-keeping places', its use for scientific purposes will depend entirely on the granting of permission by those indigenous people with responsibility for it. For this to occur, a close cooperation and mutual trust between the custodians and the scientists is essential. It is also vital that the material be treated in a manner that recognizes its cultural, spiritual, scientific and educational importance, and it should be regarded with the same degree of respect as given to any grave or burial place.

If scientific investigation is permitted, the results of the studies must be openly shared with those who have a direct cultural and/or religious interest in the material. Failure to do so reflects cultural arrogance by scientists, and is an implicit rejection of the potential significance of the remains to living cultures. It can only be hoped that, if more indigenous people become anthropologists, these barriers will begin to break down.

In the case of prehistoric remains, where no direct descendants can be established, it is our view that there should be a bias against handing the material over to indigenous groups. With no directly identifiable descendants, the interests of humanity in general should take precedence over the interests of any specific group, and the remains should be made available for reputable scientific investigation. The lack of clear associations with the living also suggests that gaining consent to undertake scientific study is of secondary importance, as the main concern is that the material is used to contribute to our understanding of human development and culture. These guidelines should apply equally to recently uncovered material from the prehistoric past. Without the activity of the scientists involved in unearthing the material, it would remain unknown, lacking any cultural, religious or scientific significance. □

D. Gareth Jones and Robyn J. Harris are in the Department of Anatomy and Structural Biology at the University of Otago, PO Box 913, Dunedin, New Zealand (e-mail: gareth.jones@stonebow.otago.ac.nz).

IMAGE
UNAVAILABLE
FOR
COPYRIGHT
REASONS

FRANK SPOONER PICTURES

What is now most important is establishing the best uses of the material to assist humanity as a whole.

Memorizing the floral ABC

Gerd Jürgens

Homeotic genes control the development of cell groups — from body segments in the fruitfly *Drosophila* to floral organs. The cloning of the *CURLY LEAF* gene in *Arabidopsis* reveals similar regulation of homeotic genes in this plant as in *Drosophila*.

Flowers and insects may have more in common than meets the eye. Both are composed of repeated units — whorls of floral organs, or body segments — which owe their specific identities to homeotic genes. Although they are unrelated by DNA sequence, the fly and flower genes have essentially the same biological effects: if they are inactivated or expressed ectopically (in the wrong place), normal structures are formed at the wrong positions — legs where antennae should be or petals in place of stamens. These homeotic transformations (making a structure look like some other structure elsewhere in the body) gave the group of developmental control genes their generic name. The similarities between the two systems have now been taken one step further by Goodrich and colleagues¹, who report on page 44 of this issue that *CURLY LEAF*, a negative regulator of an *Arabidopsis* floral homeotic gene, seems to be related to a Polycomb-group repressor of *Drosophila* homeotic genes.

Flowers have only four types of organs, which are arranged in concentric whorls (Fig. 1): leaf-like sepals in the peripheral whorl 1, showy petals in whorl 2, stamens or male reproductive organs in whorl 3, and carpels or female reproductive organs in the central whorl 4. Homeotic transformations of organs (for example, the conversion of sepals into carpels) in two distantly related plant species, *Arabidopsis* and *Antirrhinum*², indicated that simple genetic rules might underlie the construction of the flower. Indeed, these observations led to an elegant unifying concept, the ABC model³, which stipulates that three classes of homeotic gene activities (A, B and C) are sufficient to generate the four types of floral organs in a combinatorial fashion — class A acts in whorls 1 and 2, class B in whorls 2 and 3, and class C in whorls 3 and 4. So class A genes alone make sepals, A and B together make petals, B and C stamens, and class C genes alone make carpels. To account for the transformation of sepals into carpels and vice versa, it was proposed that the class A and class C genes mutually inhibit each other.

The ABC model not only correctly predicted the phenotypes of double and triple mutants, but it also withstood more sophisticated tests. For example, ectopic expression

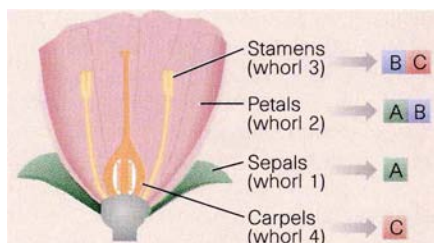


Figure 1 The activities of floral homeotic genes determine the identity of floral organs. According to the ABC model, class A genes encode sepals (whorl 1); class A and B genes together make petals (whorl 2); class B and C genes are required to produce stamens (whorl 3); and class C genes alone encode carpels (whorl 4). The class A and class C genes mutually inhibit one another. Goodrich *et al.*¹ describe a gene, *CURLY LEAF*, which seems to repress the expression of a class C gene in *Arabidopsis* outside its normal domain.

of class B genes in the absence of A-gene function transformed all floral organs into stamens⁴, as predicted from the model. It is now known that floral homeotic genes encode transcriptional regulators, all of which, with one exception, contain a MADS box (a conserved DNA-binding domain). The genes are transcribed only in those whorls that are affected by mutation in the genes themselves (again with one exception). Transcription starts early in floral-organ development, and continues during the period of growth until the cells differentiate.

How does the *CURLY LEAF* (*CLF*) gene fit into this scheme? As the name suggests, *clf* mutations lead to leaf curling, which can also be caused by ectopic expression of the class C gene *AGAMOUS* (*AG*)⁵, or the class B genes⁴ *APETALA3* (*AP3*) and *PISTILLATA* together. Goodrich *et al.*¹ have found that in *clf* mutants, *AG*, and to a lesser extent *AP3*, is indeed expressed in leaves, indicating that the normal function of *CLF* is to repress *AG*. The outer whorls of *clf* flowers are also abnormal, and *AG* transcripts accumulate in petals during the later stages of flower development (although the initial expression of *AG* is normal).

What, then, is the developmental significance of the *CLF* gene? A comparison with *Drosophila* development might give a clue.

Here, homeotic selector genes are activated in the primordia that give rise to specific segments, in response to transient signals from segmentation genes. The initial pattern of active and inactive genes is subsequently maintained through many rounds of cell division⁶. The maintenance (or memory) system involves Polycomb-group genes, which are involved in continuous repression. These genes were named after the dominant mutant-phenotype of the founder gene of the group — male *Drosophila* not only have sex combs on their forelegs, but also on their midlegs and hindlegs. Most of the Polycomb-group genes that have been analysed so far encode structurally diverse chromosomal proteins⁷. Mutations in Polycomb-group genes result in ectopic expression of homeotic selector genes, for example, the Bithorax-cluster genes; however, this only occurs at a later stage of embryogenesis.

The functional similarities indicate that the *CLF* gene and the Polycomb-group genes may have comparable functions in maintaining the inactive state of homeotic genes. The startling observation made by Goodrich *et al.*¹ is that the deduced *CLF* protein sequence shares extensive stretches of amino acids — such as the functionally uncharacterized SET domain, and several cysteine-rich domains — with the chromosomal protein Enhancer of zeste⁸, which is a *Drosophila* Polycomb-group member.

The new results raise several questions. Is *CLF* just the first representative of a group of functionally, but not necessarily structurally, related maintenance genes in plants? If this were the case, would double mutants show a more complete transformation of homeotic units, as Polycomb-group double mutants do in *Drosophila*? And are other aspects of plant development affected? The *clf* mutant could also be used to address a long-standing question in plant development. Plant cells have long been viewed as being flexible in their developmental fates, and subject to continuous signals from their surroundings. In contrast, stable expression of homeotic gene activities in *Drosophila* confers identity in a cell-autonomous fashion from early in development. These two views are not necessarily mutually exclusive, and generating clones of mutant *clf* cells might be a means of assaying for autonomous cell-fate decisions in plant development. □

Gerd Jürgens is in the Lehrstuhl für Entwicklungs-genetik, University of Tübingen, Spemannstrasse 37–39, D-72076 Tübingen, Germany.

- Goodrich, J. *et al.* *Nature* **386**, 44–51 (1997).
- Coen, E. S. & Meyerowitz, E. M. *Nature* **353**, 31–37 (1991).
- Weigel, D. & Meyerowitz, E. M. *Cell* **78**, 203–209 (1994).
- Krizek, B. A. & Meyerowitz, E. M. *Development* **122**, 11–22 (1996).
- Mizukami, Y. & Ma, H. *Cell* **71**, 119–131 (1992).
- Lawrence, P. A. & Morata, G. *Cell* **78**, 181–189 (1994).
- Kennison, J. A. *Annu. Rev. Genet.* **29**, 289–303 (1995).
- Carrington, E. A. & Jones, R. S. *Development* **122**, 4073–4083 (1996).

Growing up in a two-parent family?

Jack J. Lissauer

The nearly planar and circular orbits of the planets in our Solar System led two eighteenth-century natural philosophers, Kant and Laplace, to postulate that planets form in circumstellar disks. So there was great excitement when a dust disk was finally seen¹ around the nearby bright star β Pictoris in 1984. The study of this disk has yielded tremendous scientific returns over the past decade. Images of similar disks around other stars would give us clues about the diversity of planetary systems, but until now they have been elusive². Therefore the image of what may be a dust disk around the binary stars BD+31°643 presented by Kalas and Jewitt on page 52 of this issue³ is of great interest.

In most simulations, circumstellar disks of gas and dust are a natural by-product of star formation from the collapse of rotating molecular-cloud cores^{4,5}. They appear to be quite common around young stars that are still contracting towards the main sequence^{6,7} (the main sequence is the longest phase of a star's luminous lifetime, during which hydrogen fusion in the core produces enough energy to stave off further gravitational contraction). Also, many young stars have spectra much broader than a black body at a single temperature, implying that each source radiates over a wide range of temperatures, perhaps from different regions of a

circumstellar disk⁸. The fraction of stars with such infrared excesses decreases with stellar age, suggesting that protoplanetary disks around solar-mass stars last between one million and thirty million years⁹. Older, main-sequence stars generally show little evidence for circumstellar material.

Nevertheless, the Infrared Astronomy Satellite (IRAS) found excess mid- to far-infrared radiation coming from the environs of the main-sequence stars Vega and β Pictoris, and several others¹⁰. Small infrared excesses around main sequence stars are now known to be fairly common. β Pic's infrared excess is exceptionally large^{11,12} for a main-sequence star, and, soon after the IRAS measurements, Smith and Terrile found out why¹. They used a coronagraph, which blocks out the otherwise overwhelming light of the star, to get a near-infrared image of what appears to be a nearly edge-on disk around β Pictoris. The disk extends beyond 1,000 astronomical units from the star (1 AU is the Earth–Sun distance), and so it is much larger than our planetary system. It has been intensively studied since then^{13,14}, and the returns have been commensurate with the effort¹⁵.

The region near the star contains little dust — perhaps it has accreted to form planets. The amount of dust reaches a maximum around 10–20 AU from β Pic, decreases slowly

ly out to 100 AU and then more rapidly until it fades into the sky background. Spectra of the disk show the presence of silicates¹³. Temporally varying, redshifted absorption lines in the star's spectrum suggest that comets are quite abundant in this system¹⁴.

The study of disks around other main sequence stars would enable us to place the β Pic disk in context and so learn a lot more about the late phases of planetary growth. But despite many attempts², none have been found. So the new image³ of what appears to be a dust disk around the binary main sequence stars BD+31°643 may turn out to be very important.

This new candidate looks broadly like the β Pic disk, but it differs in many respects. It orbits a pair of stars, both of which are more massive than β Pic (five solar masses, compared with 1.5) and probably much younger. The environment is much dustier, as shown by nearby nebulosity that is illuminated by the stars at the centre of the disk (Fig. 1). It is several times as large as β Pic's disk, and the lower bound that can be placed on its mass is about 100 times as great. (In both disks, bodies ranging in size from ball bearings to small stars could contain substantially more mass, yet remain unseen because of their small total surface area.)

Another, probably even more significant difference is the blue colour of the disk around BD+31°643, which contrasts sharply with the neutral colour of the β Pic disk. Colour variations in astrophysical dust are mostly due to differences in size, because particles are progressively less effective at scattering photons of longer wavelength than their diameters. The blueness of the particles around BD+31°643 implies a typical diameter of only 0.1 μ m, a tenth of the size of particles in the β Pic disk.

The two stars of BD+31°643 are bright, and for 0.1- μ m grains the outward pressure exerted by radiation substantially exceeds the gravitational pull. So dust present at the time of star formation would not still be this close to the stars. Dust formed recently in a cooling bipolar flow (a feature often observed around very young pre-main-sequence stars) might, but the image looks more like a disk than a bipolar outflow. Kalas and Jewitt propose instead that the observed dust is continuously resupplied by the erosion of larger solid bodies — known as planetesimals, because they could serve as building blocks for planets — that accreted when the disk was younger and more massive.

If this model is correct, it would imply that planetesimals, and quite possibly planets, can form about massive binary stars. But the similarity in colour between the disk and the nearby nebulosity suggests a simpler alternative: the small grains in the disk may be mixed with and held in orbit by a much larger quantity of gas which circles the star on nearly keplerian

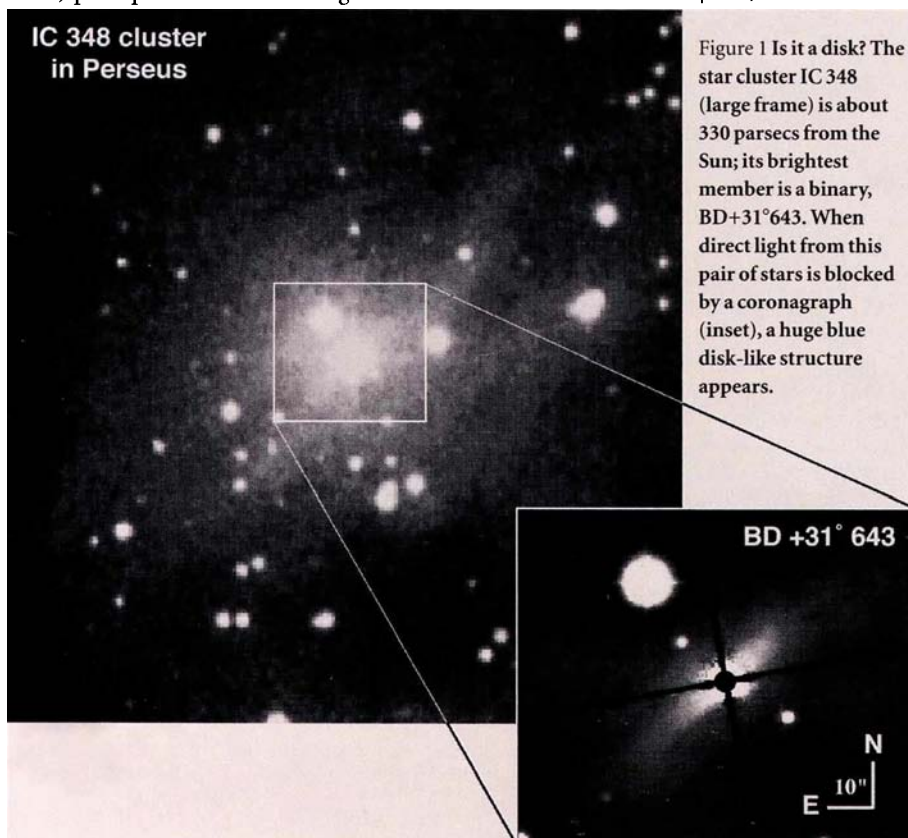


Figure 1 Is it a disk? The star cluster IC 348 (large frame) is about 330 parsecs from the Sun; its brightest member is a binary, BD+31°643. When direct light from this pair of stars is blocked by a coronagraph (inset), a huge blue disk-like structure appears.

orbits. These stars, being so massive, have reached the main sequence quickly, and the system is only about a million years old (comparable to many presumably protoplanetary disks surrounding less massive pre-main-sequence stars), so the gas may still be around. The presence of this dust could then lead to the opposite conclusion: that accretion of planetesimals within such an environment is a slow and/or inefficient process.

So even if the structure detected by Kalas and Jewitt is indeed a circumstellar disk, further research will be required to determine its implications for models of planetary growth. Nonetheless, if β Pic is any indication, data from BD+31°643 over the next decade may well provide us a wealth of new information about planet formation around massive binary stars. □

Jack J. Lissauer is in the Space Science Division, NASA Ames Research Center, Moffett Field, California 94035, USA.

1. Smith, B. A. & Terrile, R. J. *Science* **226**, 1421–1424 (1984).
2. Smith, B. A., Fountain, J. W. & Terrile, R. J. *Astron. Astrophys.* **261**, 499 (1992).
3. Kalas, P. & Jewitt, D. *Nature* **386**, 52–54 (1997).
4. Cameron, A. G. W. *Icarus* **1**, 13–69 (1962).
5. Cassen, P. M. & Moosman, A. *Icarus* **48**, 353–376 (1981).
6. McCaughrean, M. J. & O'Dell, C. R. *Astron. J.* **111**, 1977–1986 (1996).
7. Beckwith, S. V. W. & Sargent, A. I. *Nature* **383**, 139–144 (1996).
8. Adams, F. C., Lada, C. J. & Shu, F. H. *Astrophys. J.* **326**, 865–883 (1988).
9. Strom, S. E., Edwards, S. & Skrutskie, M. F. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 837–866 (Univ. Arizona Press, Tucson, 1993).
10. Gillett, F. C. in *Light on Dark Matter* (ed. Israel, F. P.) 61–69 (Reidel, Dordrecht, 1986).
11. Backman, D. E. & Paresce, F. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 1253–1304 (Univ. Arizona Press, Tucson, 1993).
12. Lissauer, J. J. in *Circumstellar Dust Disks and Planet Formation* (eds Ferlet, R. & Vidal-Madjar, A.) 327–331 (Editions Frontiers, Gif sur Yvette, 1994).
13. Telesco, C. M. & Knacke, R. F. *Astrophys. J.* **372**, L29–L31 (1991).
14. Vidal-Madjar, A. et al. *Astron. Astrophys.* **290**, 245–258 (1994).
15. Ferlet, R. & Vidal-Madjar, A. (eds) *Circumstellar Dust Disks and Planet Formation* (Editions Frontiers, Gif sur Yvette, 1994).

Visual perception

Predicting the present

Patrick Cavanagh

Richard Gregory has said that vision allows us to live a little bit in the future — to react to things before they happen. But in one sense, he is quite wrong. There is a delay of about 50 msec between the arrival of photons at the retina and the response of higher visual areas. During this time a moving object has continued on its way, so the higher visual areas will be analysing the past history of the object, not its current location. On page 66 of this issue, Nijhawan¹ brightens our outlook by showing that in some cases we can regain at least the present: we can see things not as they were, but as they probably are. This kind of predictive perception, compensating for neural delays, can only happen for objects that are undergoing smooth change (in location, brightness or size, for example), and visual motion is the basis of this particular example.

Suppose the brain does extrapolate the perceived location of a smoothly moving bar ahead of its actual position, potentially compensating for neural delays². What should happen then, if a thin target line is flashed briefly, exactly on top of the moving bar? The flashed line will be perceived at its actual location but, according to Nijhawan, the position at which the moving bar is seen will be extrapolated to a location ahead of the target line (Fig. 1). Moreover, if the moving bar has one colour and the flashed target has a different colour, the stimulus on the retina will have the sum of the two colours. Yet when they are disentangled, the bar and the target line are sorted out into their original colours — the bar is green, and the flashed target is red. So although the stimulus is yellow

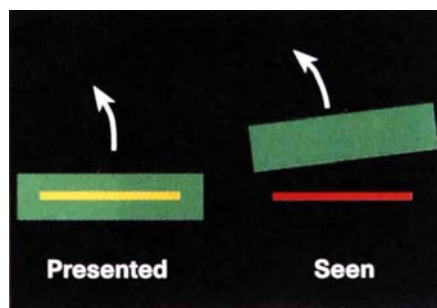


Figure 1 Neural delays that occur between the arrival of photons at the retina and the responses of higher visual areas mean that we live forever slightly in the past. But Nijhawan¹ has used a simple technique to show that the brain can, under certain circumstances, compensate for the neural delay. The red target line is flashed briefly at the centre of the smoothly moving green bar. Nevertheless, the green bar is seen ahead of the line at the moment the flash is perceived. The location of the moving bar has been extrapolated along the direction of motion, and the appropriate colour of the flashed line (red) has been recovered from the sum present in the physical stimulus on the retina (yellow).

low (the sum of the two), the observer sees a red flash trailing a green bar. How has this decomposition occurred?

Much of our understanding of the analysis of colour in the visual system emerged from observing the consequences of mixing coloured lights — discovering how, for example, red and green combine to provoke the experience of yellow. In reversing this process, the new study shows that the green

colour of the bar can be subtracted from the yellow to yield the original red of the target. There are other instances where one colour can be discounted, revealing another. For example, viewing a multi-coloured Mondrian-like pattern in, say, a reddish light leads to correction for the illumination. So patches whose spectra would be seen as yellow if viewed in isolation, seem close to the green that would be seen under white light. Nijhawan's example may involve a similar mechanism of discounting or subtraction but, unusually in his case, the two colour components are also seen as being spatially separated.

The use of yellow as the combination colour is probably of no special importance — a red herring so to speak. Certainly, other colour pairs ought to be equally well decomposed, for example, yellow and blue, or white and green. But what about black? Imagine that the target is a dark line, briefly presented as an area of decreased brightness centred in the moving green bar. When the target is seen trailing the green bar, it ought to seem blacker than black; a darker area in a background that is already black. Or if the moving bar is textured and the target is briefly presented as a filled, untextured region, subtraction would predict that, when it is seen alone, the target line will appear as the complementary texture to the moving bar.

The second component of Nijhawan's effect, extrapolation, is both more striking and more controversial than subtraction. Neural delays between the arrival of photons at the retina and the responses of higher visual areas mean that it is the past history of an object that is analysed, not its current state. The author claims that this lag is corrected — the location at which the object is perceived is extrapolated ahead, along its path, to compensate for the delay. But why bother? Neural delays affect all of our sensations: our perception of auditory events must lag behind the arrival of sound at the ear; and our awareness of our moving arms and legs will lag behind the proprioceptive and visual cues of their position. The benefit of using extrapolation — predictive perception — in all of these modalities is not immediately obvious. Moreover, the cost is that signals undergoing change which will permit extrapolation should be mislocalized relative to signals which have abrupt onsets or offsets. For example, when two objects are moving towards each other at a constant velocity, the sound of their collision should seem to lag behind the visual perception of the collision. (Observers, however, are not very sensitive to small asynchronies between modalities such as hearing and seeing³.)

How would extrapolation occur? A sensed location needs to be assigned to individual neurons, or groups of neurons, and that location is linked in some way⁴ to the areas of the retina that activate the neurons.

In the case of a motion-sensitive unit, Nijhawan's results indicate that the location is simply assigned to a region that is shifted away from the receptive-field centre of the neuron, in the direction of the preferred motion.

There is an alternative to extrapolation which could also explain the striking results — persistence. A stimulus remains visible for a brief moment after it has physically been turned off, and this persistence is shorter for a moving stimulus than it is for a flashed stimulus⁵. This difference acts as a 'motion deblurring' mechanism, to remove the smeared traces which should otherwise trail moving objects. A short while after the moving bar has passed the flashed target line, no trace of the bar remains there, whereas the flashed target lives on longer (persists), to be seen as physically separate from the moving bar. Nijhawan has addressed this possibility in his last experiment, but two other considerations argue that persistence is not the sole cause of the phenomenon.

First, the motion-specific suppression would also have to be colour-specific; it must extinguish only the green bar, and not the red line which falls at the same location. But colour-specific motion mechanisms⁶ seem to act sluggishly^{7–9}, and it is unlikely that they would be fast enough to support deblurring. Second, the trailing suppression can speed the disappearance of the moving bar's trace, but it should not affect the bar's perceived time of arrival at the location of the target. As the arrival of the moving bar and the onset of the flashed line physically occur at the same time, the bar and the line should seem, at least initially, to overlap. The longer persistence of the flashed line may then give the impression that it remains on after the moving bar has passed over, but persistence cannot explain how the bar and line can initially appear with no overlap at all.

Overall, Nijhawan's observation presents two visual mechanisms that are more capable than we might have imagined: a subtraction process which can disentangle the combined colours of superimposed stimuli and reassign them to different locations in space; and an extrapolation process which allows us to live, if not a little bit in the future, at least in the present. □

Patrick Cavanagh is in the Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, USA.

1. Nijhawan, R. *Nature* **386**, 66–69 (1997).
2. Nijhawan, R. *Nature* **370**, 256–257 (1994).
3. Sekuler, R., Sekuler, A. B. & Lau, R. *Nature* **385**, 308 (1997).
4. Ahumada Jr, A. J. in *Computational Models of Visual Processing* (eds Landy, M. S. & Movshon, J. A.) 23–34 (MIT Press, Cambridge, MA, 1991).
5. Burr, D. *Nature* **284**, 164–165 (1980).
6. Cropper, S. J. & Derrington, A. M. *Nature* **379**, 72–74 (1996).
7. Cavanagh, P., Tyler, C. W. & Favreau, O. E. *J. Opt. Soc. Am. A* **1**, 893–899 (1984).
8. Derrington, A. M. & Henning, G. B. *Vision Res.* **33**, 799–811 (1993).
9. Gegenfurtner, K. R. & Hawken, M. J. *Vision Res.* **35**, 1547–1563 (1995).

Photochemistry

Very small arrays

Devens Gust

Visitors to the Plains of San Agustin in central New Mexico are captivated by the immense antennas of the Very Large Array radio telescope. Much more abundant on these plains, and no less fascinating, are the antennas of photosynthetic organisms, which gather light from the Sun. Artificial antennas that do the same job could revolutionize solar energy technology; and two papers^{1,2} published late last year in the *Journal of the American Chemical Society* describe chemical structures that funnel energy in the right way.

Plants use complex photosynthetic reaction centres to convert energy from sunlight into electrochemical potential. But although these reaction centres are superb energy transducers, they are metabolically expensive to manufacture and they do not capture light well at all wavelengths. Photosynthetic organisms have solved these problems by developing efficient antennas that absorb light throughout the visible spectrum, and transport their excitation to the reaction centres, where it is used to do photosynthetic work.

Photochemists have long dreamed of harvesting the Sun's energy by mimicking the photosynthetic process, and already there are artificial reaction centres whose performance rivals that of nature^{3,4}. So some chemists have now turned their attention to making artificial antennas modelled on natural ones. The two new approaches are both based on covalently bonded organic pigments that absorb light and funnel excitation energy to a central chromophore. In one, a dendritic array of phenylacetylene chromophores absorbs ultraviolet light and transports the resulting excitation to a perylene moiety¹. In the second, porphyrin molecules, which absorb visible light, have been linked via ethyne bridges to form linear and two-dimensional light-harvesting arrays².

Photosynthetic organisms live in environments with different qualities of light, and so have evolved a variety of antennas to adapt to these conditions⁵. In general they use arrays of pigments of different colours, such as chlorophylls, carotenes and bilins. The colour of each cluster or 'pool' of pigments is determined by the energy of an electronic excited state, and careful spatial arrangement of these energy levels is what produces the energy-funnelling behaviour.

The highest-energy pool is most distant from the reaction centre. Absorbed light generates an electronic excitation that migrates from pigment to pigment within the pool, and is eventually transferred to a second pool of lower-energy pigments that is closer to the reaction centre. The excitation

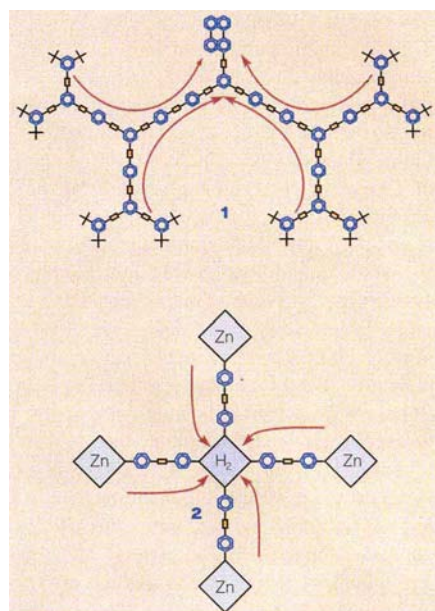


Figure 1 Two synthetic light-harvesting antennas: a dendritic phenylacetylene structure with a perylene core (1), and four zinc-containing tetra-arylporphyrins linked to a central, unmetallated porphyrin through ethyne bonds (2). The arrows show the direction in which singlet excitation energy is transferred.

continues to migrate through even lower-energy pigment pools until it reaches the reaction centre, where it is converted to electrochemical potential in the form of charge separation between two chlorophyll derivatives. This entire process occurs in a few times 10^{-11} seconds or less, with a quantum yield of nearly one.

Photosynthetic antennas are designed to direct the migration of excitation much as some multi-storey shopping malls are designed to guide patrons. A shopper enters at an upper level and wanders randomly from shop to shop within that level. Happening upon a staircase, he descends to the next level and visits the shops on that floor. Our shopper descends floor by floor, ultimately arriving at street level, where he finds his way out of the complex.

When the typical shopper leaves the mall, he finds his wallet significantly lighter than when he entered. Similarly, photosynthetic antennas extract an energy toll each time the excitation moves to a longer-wavelength pigment pool. In exchange for this loss, the excitation is transported to the vicinity of the reaction centre with a very high quantum yield, and arrives in a form that can be efficiently coupled to the centre's pigments. Funnel-type antennas of this sort are best characterized in the cyanobacteria (blue-



100 YEARS AGO

The various methods for producing photographic pictures in colour were described by Sir Henry Trueman Wood at the Society of Arts... The main object of the paper was to bring before the Society M. Chassagne's promising process for the photographic reproduction of colour, but the opportunity was taken to summarise the whole question of colour-photography. Though the *rationale* of the new process remains a mystery, there can be no question that very remarkable results are produced. Even more striking than M. Chassagne's pictures, however, are some transparencies exhibited at the same meeting by Mr Bennetto, of Newquay, in Cornwall... In strictness, of course, it is not possible to know for certain that a particular result is produced by a particular process unless the nature of that process is also known; and from that point of view, it is perhaps allowable to regard Mr Bennetto's pictures with some degree of philosophical suspicion. But he has been put to tests which it is difficult to suppose he could have satisfied did he not in fact do what he claims to do. He was requested to focus his apparatus on an easel. When he had done so, he was blindfolded, and on the easel was placed an impossible picture, painted in impossible colours, which he had never seen before. This he photographed and developed, still without seeing the original, with the result that the impossible colours were reproduced in the photograph obtained.

From *Nature* 4 March 1897.

50 YEARS AGO

It is thus to a world of individualist enterprise that Dr Bush makes his appeal for the establishment of a council to control the scientific policy of the United States. He may, of course, be wrong in assuming that advances in science when put to practical use necessarily "mean more jobs, higher wages, shorter hours... and higher standards of living" — since these things depend at least as much on socio-economic factors as on scientific progress. But whether the atmosphere be the rich and individualist one of the United States, or the Spartan and socialist one of Great Britain to-day, it seems to be generally recognized that the time has come for the State to play a central part in stimulating scientific progress.

From *Nature* 8 March 1947.

green algae), purple bacteria and green photosynthetic bacteria.

The synthetic antenna (1 in Fig. 1) described by Moore and co-workers¹ is also a dendritic molecule — a series of branched, repeating units that originate from a central core. At the heart of the antenna is a perylene pigment with a relatively low-energy first excited singlet state. This is bonded to a tree of phenylacetylene chromophores whose excited-state energies increase as one moves towards the edge. Light is absorbed by the phenylacetylene units and, within a few picoseconds, excitation is funnelled down the energy gradient to the perylene, with a quantum yield of about 0.98. The absorption cross-section of the perylene group — its ability to capture nearby photons — has been increased enormously.

Porphyrins, which are structurally similar to the chlorophylls, absorb visible light strongly. The linear and branched antenna arrays described by Lindsey and colleagues² are made from porphyrins linked by rigid ethyne groups. One example is the pentameric structure (2 in Fig. 1), consisting of four peripheral porphyrins each containing a zinc atom, bonded to a fifth, zinc-free porphyrin at the centre. The chemical bonds are

designed so that each porphyrin interacts only weakly with its neighbours, and so retains its spectroscopic identity. The Lindsey group has also reported a linear 'photon-wire' that operates on similar principles⁶.

Synthetic antennas based on directional energy funnelling will eventually be married to artificial photosynthetic reaction centres, bringing us one step closer to mimicking the entire photosynthetic energy-conversion process in the laboratory. The new antennas may be used in molecular-scale solar energy transducers, optical sensors and probes, and optoelectronic devices. Perhaps the familiar cityscape of microchips will soon be graced by a few trees. □

Devens Gust is in the Center for the Study of Early Events in Photosynthesis, Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona 85287, USA.

1. Devadoss, C., Bharathi, P. & Moore, J. S. *J. Am. Chem. Soc.* **118**, 9635–9644 (1996).
2. Wagner, R. W., Johnson, T. E. & Lindsey, J. S. *J. Am. Chem. Soc.* **118**, 11166–11180 (1996).
3. Gust, D., Moore, T. A. & Moore, A. L. *Acc. Chem. Res.* **26**, 198–205 (1993).
4. Wasielewski, M. R. *Chem. Rev.* **92**, 435–461 (1992).
5. van Grondelle, R., Dekker, J. P., Gillbro, T. & Sundström, V. *Biochim. Biophys. Acta* **1187**, 1–65 (1994).
6. Wagner, R. W. & Lindsey, J. S. *J. Am. Chem. Soc.* **116**, 9759–9760 (1994).

Materials science

The how and why of thermal contraction

Robert W. Cahn

Most solids expand when heated; this is a simple consequence of the enhanced amplitude of thermal vibration at higher temperature. Any instance of a material that contracts or remains invariant, instead of expanding, requires some special explanation. Always, the unavoidable lengthening of interatomic bonds when heated has to be counteracted by some other physical factor. A recent paper¹, by a group headed by Arthur Sleight, reports on a very detailed study of zirconium and hafnium tungstates, two isomorphous crystal species both of which contract when heated, and the investigators succeed in teasing out the strange physical origin of this behaviour.

This study is an impressive illustration of the insights that can be garnered from a very thorough X-ray analysis of a crystal structure, at both ambient and high temperatures, and carefully estimating the degree of precision of atomic positions as well as unravelling the anisotropic amplitude of thermal vibrations of individual atoms. Here, X-ray analysis has been complemented by various spectrographic studies. The structure consists of ZrO_6 (or HfO_6) octahedra linked through oxygen atoms to WO_4 tetrahedra, leaving one 'terminal' (unlinked) oxygen on each tetrahedron.

The authors show that the constituent polyhedra librate (tilt) in a coupled manner; this involves transverse vibrations of Zr–O–W bonds and a shortening of the 'non-bonded' W–O pairs. The upshot of the transverse vibrations is an effective shortening of Zr–O–W bonds and thus an overall contraction. A framework lattice of strong metal–oxygen bonds with two-coordinated bridging oxygens seems to be an essential precondition of thermal contraction in oxides.

The previous year, another group of investigators led by Sleight² examined a group of crystals in the $ZrV_{2-x}P_xO_7$ solid-solution series, most of which also show thermal contraction over part of the temperature range examined. These crystals again consist of linked polyhedra. As Sleight explains in a beautifully clear overview paper³ on thermal contraction, the structure 'wants' to reduce its symmetry on cooling and thereby bend the originally straight metal–oxygen–metal bonds, but (because of the structural complexity) "nature is unable to find a suitable transition to a lower symmetry where the M–O–M bonds can all bend statically away from 180°. The best it can accomplish is a structure where 89% of these angles have bent away from 180°". It is suggested that

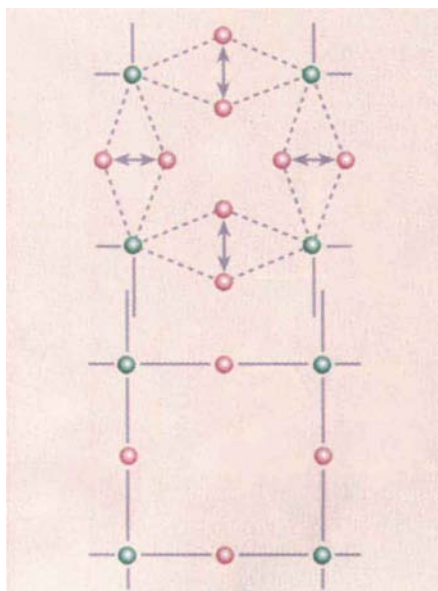


Figure 1 Contraction caused by increased thermal motion. Green circles are cations and red circles are oxygen. As oxygens vibrate normal to the line of the bonds, cations are pulled closer together³.

these crystals are in a metastable condition; they want to transform to a well defined low-symmetry structure on cooling but are structurally 'frustrated' from doing so. The 'static' bending would interfere with the 'dynamic' (thermal) vibration of the oxygen atoms (Fig. 1) that reduces the metal-metal distance and thus causes thermal contraction in this kind of crystal. Thus, chemical frustration, preventing some bonds from being statically bent, may be a further precondition of strong thermal contraction.

An important feature of these various thermally contracting crystals is that they are crystallographically cubic, and the change of dimensions with rising temperature is therefore isotropic. This means that, in a polycrystal, no internal stresses are generated by non-matching shape changes of neighbouring grains. In such brittle materials, anisotropic thermal expansion (or contraction) is apt to lead to micro-cracking⁴, and this limits the value of existing low- or zero-expansion non-cubic ceramics such as β -eucryptite (LiAlSiO_4), cordierite ($\text{Mg}_2\text{Al}_2\text{Si}_2\text{O}_{10}$) and β -spodumene ($\text{Li}_2\text{Al}_2\text{O}_4 \cdot n\text{SiO}_2$), 'glass-ceramic' phases which are commercially made by crystallizing glasses, for applications such as astronomical telescope mirrors and thermal-shock-proof cookware. Sleight³ explains that in these anisotropic silicates, weak Li-O bonds expand and thereby force the very strong Al-O and Si-O bonds, which do not lengthen with rising temperature, to pucker, and thus cause just one crystal axis to contract thermally.

Sleight has patented some of his new isotropic thermally contracting crystal species⁵, and newspaper comment has indi-

cated that there is great commercial interest in them — as components of dental fillings that expand on cooling, for example. If these special materials are mixed with normal isotropic ceramics, it should be easy to generate zero thermal expansion combined with no tendency to microcracking, and this may yield alternative candidates for such high-profile uses as telescope mirrors.

The oxides discussed up to now are not the only category of solids with low, zero or negative thermal expansion. A number of metals and alloys share these characteristics. Historically the oldest are the invars, a term originally applied to $\text{Fe}_{65}\text{Ni}_{35}$. As this alloy is cooled, it changes from paramagnetic to ferromagnetic a little above room temperature; below this Curie temperature, there is an exceptionally large 'spontaneous magnetostrictive volume expansion' over a small temperature range that compensates for the normal thermal contraction on cooling, so that, around ambient temperature, there is approximately zero thermal expansion. "Invar" is in fact a trademark of Imphy SA in France, who have manufactured it ever since Charles-Edouard Guillaume discovered this material and its key property in 1896. Invar has long been used in the construction of instruments and devices (such as TV masks) in which key dimensions must be invariant with temperature over a limited temperature range near ambient.

A whole range of alloys (mostly inter-metallic compounds) has more recently been found to have similar characteristics^{6,7}. What are the factors that favour a magnetostrictive volume expansion large enough to secure invar-like behaviour? The magnetostrictive expansion is linked to the increase in the kinetic energy of 3d electrons in a transition metal as a result of spin polarization induced by exchange interaction between the magnetic spins as they line up on cooling. A volume expansion can reduce this kinetic energy by increasing the density of electronic energy states in the 3d bands, and its magnitude is linked with the exact band structure of a particular alloy or compound. So the origin of the anomalous thermal expansion here is totally different from that for Sleight's oxides.

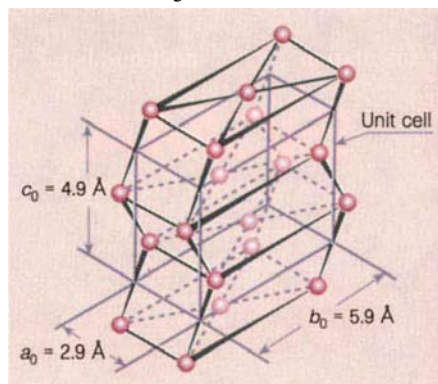


Figure 2 The crystal structure of α -uranium¹².

Negative thermal expansion has also long been known in α -uranium (the ambient-temperature, orthorhombic allotrope of metallic uranium) and in face-centred cubic δ -plutonium, one of the many allotropes of that metal, stable between 592 and 724 K. The crystal structure (Fig. 2) of α -uranium has strong, short covalent bonds within puckered sheets, and much weaker bonding between the sheets. The thermal expansion in the a and c directions is normal (positive), but in the b direction it is strongly negative⁸. With unaltered length of the strong bonds, the expansion on heating in the c direction tends to unpucker the sheets and thus draw them closer together in the b direction. This behaviour is similar to that of cordierite³, and is different from that of the invars and of δ -plutonium (and from that of the isotropic oxides).

Lander *et al.*⁹, in a recent magisterial overview of the solid-state properties of uranium, comment that once the thermal expansion anisotropy had been shown to be intrinsically linked with the inconvenient behaviour of the metal under irradiation, the metal was promptly abandoned as a nuclear fuel (though not quite in all reactors) and replaced by isotropic uranium dioxide. As they say: "This surely represents one of the most rapid changes of technology driven by basic research".

In contrast, the thermal contraction of δ -plutonium has been much debated, but all the explanations relate to the electronic band structure. Thermal contraction results if the density of states at the Fermi energy changes sufficiently sharply with energy¹⁰, particularly so if there are large numbers of overlapping bands. Or it may be that the peculiarities of this phase are attributable to the f band being exceptionally narrow¹¹. But the fact that δ -plutonium contracts thermally whereas all the other allotropes expand normally shows, if proof were needed, that thermal contraction is a function of crystal structure and not of the chemical constitution of a material alone. □

Robert W. Cahn is in the Department of Materials Science and Metallurgy, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, UK.

1. Evans, J. S. O., Mary, T. A., Vogt, T., Subramanian, M. A. & Sleight, A. W. *Chem. Mater.* **8**, 2809–2823 (1996).
2. Korthuis, V. *et al. Chem. Mater.* **7**, 412–417 (1995).
3. Sleight, A. W. *Endeavour* **19**, 64–68 (1995).
4. Taylor, D. in *Encyclopedia of Advanced Materials* (eds Bloor, D., Brook, R. J., Flemings, M. C. & Mahajan, S.) 2818–2822 (Pergamon, Oxford, 1994).
5. Sleight, A. W. *Negative Thermal Expansion Materials* US Patent 5,322,559 (1994).
6. Shiga, M. *Curr. Opin. Solid State Mater. Sci.* **1**, 340–348 (1995).
7. Andreev, A. V., de Boer, F. R., Jacobs, T. H. & Buschow, K. H. J. *J. Less-Common Metals* **118**, 349–353 (1986).
8. Lloyd, L. T. & Barrett, C. S. *J. Nucl. Mater.* **18**, 1206 (1966).
9. Lander, G. H., Fisher, E. S. & Bader, S. D. *Adv. Phys.* **43**, 1–111 (1994).
10. Varley, J. H. O. *Proc. R. Soc. Lond. A* **237**, 423 (1956).
11. Kmetko, E. A. in *Plutonium 1965* (eds Kay, A. E. & Waldron, M. B.) 222–243 (Chapman & Hall, London, 1967).
12. Holden, A. N. in *The Physical Metallurgy of Uranium 27* (Addison-Wesley, Reading, PA, 1958).

Mice perform a human repertoire

Michael Neuberger and Marianne Brüggemann

Ever since the first mouse monoclonal antibodies (mAbs) were created, a major challenge has been to obtain human mAbs for therapeutic use. People cannot be immunized with the desired antigen, so during the 1980s mouse mAbs were raised using conventional hybridoma technology and then 'humanized' on an individual basis — those parts of the mouse antibody that were not essential for binding to the antigen were replaced by corresponding human sequences. Although it was successful, this approach had to be carried out on a case-by-case basis and the mAbs retained some mouse sequences.

Around the start of this decade, two parallel technologies (phage display and transgenic mice) were exploited to generate repertoires of fully human mAbs from which those with specific binding sites could be selected ('binders'; Fig. 1)^{1,2}. Now, in the February issue of *Nature Genetics*, Mendez *et al.*³ describe the creation of mice in which large sections of the human immunoglobulin gene loci — transloci — have been integrated into the germ line. This work is exciting because it highlights the usefulness of the transgenic approach in generating high-affinity mAbs to specific human antigens.

Antigen-specific antibodies are isolated by the selection of specific binders from repertoires of antibodies with diverse specificities. *In vivo*, there are two distinct and sequential rounds of repertoire generation, each of which is followed by antigen-mediated selection (Fig. 1). The primary repertoire is generated by gene rearrangements involving the immunoglobulin variable (V)-region gene segments of the immunoglobulin heavy and light chains. Because this repertoire has a limited diversity, it will usually only give rise to antibodies of low affinity. But the V genes of binders that have been selected from this primary repertoire are

then subjected to a second wave of diversification (this time by somatic hypermutation) and antigen selection — this second wave is critical to obtaining high-affinity antibodies.

Selection of binders from the human primary repertoire can also be carried out *in vitro*. The rearranged V genes from peripheral blood lymphocytes are cloned into bacterial expression vectors, and binders are selected using phage-display technology⁴. The cloning process results in scrambling of the naturally occurring pairings of immunoglobulin heavy- and light-chain V regions, so a much larger primary repertoire can be produced. Consequently, antigen-mediated selection from this primary repertoire can yield high-affinity binders but, if need be, other *in vitro* strategies can be used to obtain further increases in affinity⁴.

In contrast, the transgenic approach exploits the natural diversification/selection strategies of the mouse, but has them operate on human rather than mouse immunoglobulin loci. There, they undergo the normal programme of rearrangement and hypermutation, which seems to work in a similar way in both mouse and human. Targeted disruption of the mouse immunoglobulin loci allows the human transloci to perform unfettered by competition from the endogenous loci⁵⁻⁷, and the human antibody repertoires can be sampled using conventional immunization strategies^{2,3,5-10}.

The size of the human loci — stretching over several megabases — has provided quite a challenge to transgenic antibody technology. But most of this DNA is devoted to the germline V segments, and all studies to date have been carried out with transloci that contain only a portion of the total V-gene repertoire (Fig. 2a). Early studies used miniloci, which contain only small numbers of selected V-gene segments (Fig. 2b). Such loci can, in fact, yield good antigen-specific

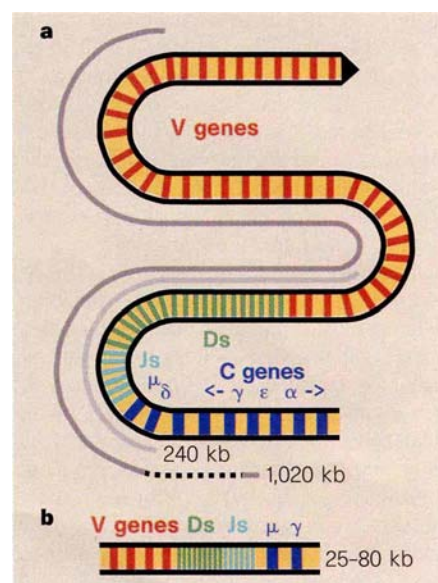


Figure 2a Yeast artificial chromosome (YAC) technology was initially used to create transloci containing only the core regions of the human immunoglobulin loci^{5,9,12} (about 240 kb in the case of the heavy-chain locus). But Mendez *et al.*³ now describe larger transloci (1,020 kb of heavy chain and 800 kb of light chain), containing a much greater proportion of the human V-gene repertoire. b, These transloci contrast with miniloci^{2,6-8,10}, which include only a restricted number of human gene segments in an artificially compact configuration.

mAbs^{7,10}, suggesting that working on a small set of V genes (but with a near-normal number of diversity (D) and joining (J) segments), junctional diversification can generate a primary repertoire that is large enough for responses to be initiated to a variety of antigens. Secondary diversification can then occur through somatic hypermutation (which is readily recruited by transloci^{8,9}).

The ability to transfer yeast artificial chromosomes (YACs) into embryonic stem cells, however, has allowed much larger segments of the human immunoglobulin loci to be introduced into the mouse germ line^{5,9,11,12}. Mendez *et al.*³ have made mice carrying some of the largest transloci to date, and they have convincingly shown the excellent performance of their translocus mice in the production of human mAbs — so there can now be little reason to prefer smaller loci. In the future, we can expect to see transloci that include the full complement of human variable- and constant-region genes. Any human antibody that is used for therapeutic purposes will probably be produced in bulk using heterologous expression systems, which will allow the isotype to be changed at will. However, the inclusion of additional constant regions on the translocus itself will allow the production of a full range of IgG antibodies in the serum of the translocus mice. It will be interesting to discover whether the presence of antigen-specific IgG

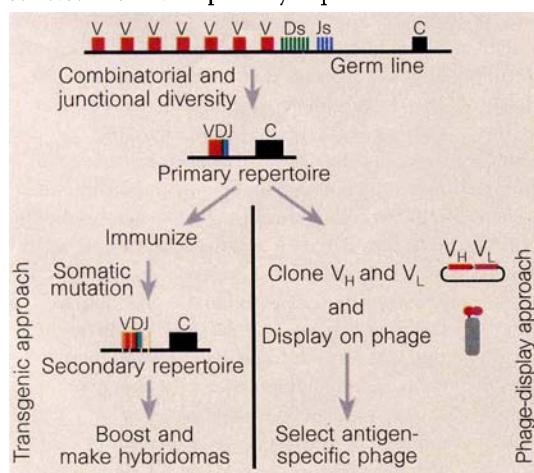


Figure 1 Strategies for making human mAbs. In the transgenic approach, mice carrying human immunoglobulin transloci are challenged with antigen: lymphocytes from the primary repertoire, which make antigen-specific antibodies, are thus subjected to somatic mutation, creating a secondary repertoire in which the immunoglobulin V genes carry nucleotide substitutions (yellow lollipops). In the phage-display approach, heavy- and light-chain V genes (V_H and V_L) are cloned into bacterial vectors and expressed on the surface of phage, and antigen-specific phage are selected *in vitro*.

(as opposed to just IgM) antibodies assists the selection of antibodies of very high affinity, a critical parameter of the technology.

From the mouse strains that have been created so far, it is clear that the transgenic approach complements phage display for the production of high-affinity human mAbs. The advantage of the transgenic approach is that, if a suitable mouse line is available, high-affinity human mAbs can readily be selected using well-established hybridoma technology. But phage display can be faster, and it is still evolving — it is probably better suited for the isolation of antibodies to evolutionarily conserved structures, with greater versatility in the nature and size of the repertoire that can be screened. Also, although the chance of obtaining high-affinity mAbs to a range of antigens is increased by having two distinct technologies available, perhaps in the future the technologies could be combined, for example by using phage display to select good binders from hyperimmunized translocus mice.

The ability to manipulate immunoglobulin YACs in yeast before introducing them into the mouse germ line could also prove useful for studying immune regulation and repertoire formation. The complex pattern

of rearrangement, expression and mutation that is shown by the immunoglobulin loci is not always mimicked by artificial miniloci; conversely, a redundancy of *cis*-regulatory elements within the natural loci may obscure their analysis by conventional gene targeting. Because of their intermediate complexity, and the ease with which they can be manipulated, YAC-based transloci may therefore not only facilitate the production of human mAbs, but they could be useful in investigating the regulation of antibody production. □

Michael Neuberger is at the Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK. Marianne Brüggemann is at the Babraham Institute, Babraham, Cambridge CB2 4AT, UK.

1. McCafferty, J., Griffiths, A. D., Winter, G. & Chiswell, D. J. *Nature* **348**, 552–554 (1990).
2. Brüggemann, M. *et al.* *Proc. Natl Acad. Sci. USA* **86**, 6709–6713 (1989).
3. Mendez, M. J. *et al.* *Nature Genet.* **15**, 146–156 (1997).
4. Hoogenboom, H. R. *Tibtech* **15**, 62–70 (1997).
5. Green, L. L. *et al.* *Nature Genet.* **7**, 13–21 (1994).
6. Lonberg, N. *et al.* *Nature* **368**, 856–859 (1994).
7. Wagner, S. D. *et al.* *Eur. J. Immunol.* **24**, 2672–2681 (1994).
8. Taylor, L. D. *et al.* *Int. Immunol.* **6**, 579–591 (1994).
9. Wagner, S. D. *et al.* *Genomics* **35**, 405–414 (1996).
10. Fishwild, D. M. *et al.* *Nature Biotechnol.* **14**, 845–851 (1996).
11. Zou, X. *et al.* *FASEB J.* **10**, 1227–1232 (1996).
12. Davies, N. P. *et al.* *Biotechnology* **11**, 911–914 (1993).

Electron cryomicroscopy

Who needs crystals anyway?

David J. De Rosier

We celebrate another success in the advancing field of electron cryomicroscopy with the publication of two papers that describe the structure of the hepatitis B virus core protein. On page 88, Böttcher *et al.*¹ report that they have used the technique to detect a new viral capsid fold — the four-helix bundle — within the virus core, at a resolution of 7.4 Å; and Conway *et al.*² (page 91) come to similar conclusions from their studies of a structure at 9 Å resolution. The work represents an important step away from the need to have ordered specimens with crystalline symmetry, and towards the structural biologist's dream of solving the atomic structures of macromolecules (and their assemblies) which do not crystallize, and which cannot therefore be studied by X-ray crystallography.

The quantitative analysis of electron micrographs began over 30 years ago, with studies on viruses. Finch and Klug³ compared shadowgraphs of plastic and wooden models with micrographs of negatively stained virus particles. Although computer-generated representations soon replaced the models, structure determination remained a trial-and-error process of adjusting a model to fit the images. Three-dimensional reconstruction eliminated trial-and-error by providing a path directly from the images to a

three-dimensional map of the structure⁴. This procedure is carried out by a computer — many digitized electron microscopic images corresponding to different views of a structure are combined to generate a three-dimensional map of the structure. Together with advances in electron microscopy, such as glucose-embedding (and, more recently, ice embedding and cryomicroscopy) and low-dose techniques that preserve the structure of the protein, image reconstruction allowed researchers to look directly at biological structures, rather than simply studying imprints made in negative stain.

In 1975, electron microscopy was used to visualize, for the first time, the α -helical bundles in the membrane protein bacteriorhodopsin⁵. The subunits of bacteriorhodopsin aggregate in the membrane into crystalline arrays. These two-dimensional crystals are good specimens from which to obtain high-resolution data, because each image contains thousands of identical units in an identical orientation. Structural analysis by electron cryomicroscopy of two-dimensional crystals now extends to essentially atomic resolution, of 3 to 4 Å.

More recently, electron microscopy was used to solve, at 9 Å resolution, the structures of a different class of molecules — helically symmetrical assemblies^{6–9}. Images of

these usually have fewer subunits than a single, two-dimensional crystal, so more images are needed to achieve the same resolution. Now, the new studies add a third class to this list — single particles that have icosahedral symmetry. Because there are still fewer equivalent units per particle, even more images are needed to achieve this resolution: Böttcher *et al.*¹ combined 6,400 images, and Conway *et al.*² combined 600.

Why is it important to solve the structure of a virus to a resolution of up to 7.4 Å, when viral structures have already been seen to atomic resolution by X-ray crystallography? The reason is that the electron microscope is able to look at structures that do not crystallize, and so cannot be analysed by X-ray crystallography. For example, an atomic model for a virus–receptor complex was derived by docking the atomic structures for the virus and for the receptor (both of which were solved by X-ray crystallography) into a three-dimensional map of the virus–receptor complex (which was obtained by electron microscopy)¹⁰. The higher the resolution of the three-dimensional map, the more precise is the model, so it is easier to adjust for the conformational changes that might accompany the formation of the virus–receptor complex.

The high-resolution maps that have been obtained by Böttcher *et al.*¹ and Conway *et al.*² look essentially the same. Both groups expressed the 150 amino-terminal residues of the 183-amino-acid hepatitis B core antigen, and they found that this partial protein assembles into icosahedrally symmetrical empty protein shells made up of a number of

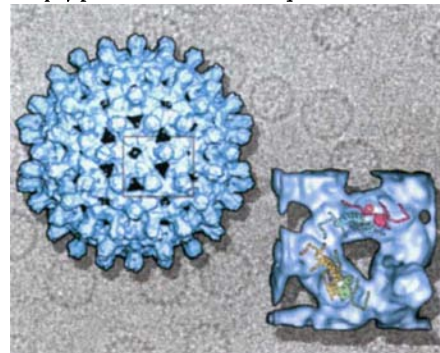


Figure 1 The ability to determine the structure of large structures at atomic resolution comes one step closer with the work of Böttcher *et al.*¹ and Conway *et al.*², who have now solved the structure of the hepatitis B virus core protein at better than 10 Å using the emerging technique of electron cryomicroscopy. The diameter of the viral capsid is 310 Å, and the boxed region has been enlarged to show the new viral fold — known as the four-helix bundle — which the technique has revealed. The four α -helices within the fold are colour coded, to show how they contribute to the bundle, and the background shows a small portion of the electron micrographs that were used to determine the structure. (Figure provided by A. C. Steven².)

clustered subunits. Each cluster contains a bundle of four α -helices, which differs from the 'jelly roll' β -barrel fold that is usually found in viral capsid proteins. Identical neighbouring polypeptide chains both contribute a pair of α -helices, in the form of a helix-turn-helix structure, to each four-helix bundle (Fig. 1).

There is less consensus about the rest of the structure. Although electron microscopy can be used to determine the handedness of molecules, neither group applied this technique to the core protein, and the handedness of the structure is not known. Indeed, the two maps seem to have opposite handedness. But there are more substantial differences still. Böttcher *et al.*¹, who used more images and achieved higher resolution, found a continuous tube of density which they suggest follows the path of the polypeptide chain. Conway *et al.*², however, could trace only three unconnected segments of density as being part of the polypeptide chain. The larger of the three, which corresponds to the four-helix bundle, seems to superpose on the trace that describes the protein fold of Böttcher *et al.*, but the two smaller segments do not. It is not surprising, at this resolution, that some parts of the polypeptide chain are difficult to trace unambiguously.

Assigning the amino-acid sequence to features in the map is also problematic. Whereas Böttcher *et al.* place the immunodominant region (the most antigenic part of the protein, at around residue 80) at the outside tip of the four-helix bundle, Conway *et al.* place it towards the base of the bundle. This issue could probably be resolved with a map of the virus decorated with antibodies that are specific to the immunodominant sequence. But whatever the final trace of the polypeptide chain, both groups can take pride in a splendid achievement.

What is next? Gazing through my non-crystal ball, I foresee two steps coming in the next few years — a 4-Å map of a helical assembly in which the peptide backbone can be traced, and a 10-Å map of an asymmetric assembly, such as the ribosome in its monomeric state. □

David J. De Rosier is at the W. M. Keck Institute of Cellular Visualization, Rosenstiel Basic Medical Sciences Research Center, Department of Biology, Brandeis University, Waltham, Massachusetts 02254-9110, USA.

1. Böttcher, B., Wynne, S. A. & Crowther, R. A. *Nature* **386**, 88–91 (1997).
2. Conway, J. F. *et al.* *Nature* **386**, 91–94 (1997).
3. Finch, J. T. & Klug, A. *J. Mol. Biol.* **13**, 1–12 (1965).
4. De Rosier, D. J. & Klug, A. *Nature* **217**, 130–134 (1968).
5. Henderson, R. & Unwin, P. N. *Nature* **257**, 28–32 (1975).
6. Jeng, T. W., Crowther, R. A., Stubbs, G. & Chiu, W. *J. Mol. Biol.* **205**, 251–257 (1989).
7. Mimori, Y. *et al.* *J. Mol. Biol.* **249**, 69–87 (1995).
8. Morgan, D. G., Owen, C., Melanson, L. A. & De Rosier, D. J. *J. Mol. Biol.* **249**, 88–110 (1995).
9. Unwin, N. *J. Mol. Biol.* **229**, 1101–1124 (1993).
10. Olson, N. H. *et al.* *Proc. Natl Acad. Sci. USA* **90**, 507–511 (1993).

Gravitational lenses

The uses of strong and weak lenses

Joachim Wambsganss

The Universe acts as a complicated lens. Matter magnifies and distorts the shapes and brightnesses of distant galaxies and quasars by the gravitational attraction of light. Fritz Zwicky predicted the use of gravitational lenses in solving cosmological problems 60 years ago¹, and the latest results presented at a meeting near Munich* show that, 18 years after the discovery of the first lens², the field has started to fulfil its promise. Large CCD cameras and new theoretical tools have allowed measurements of the galaxy-cluster mass distribution, distances to galaxies too faint for direct measurement, and the 'cosmic shear' due to large-scale structure.

Clusters of galaxies are the largest gravitationally bound structures in the Universe. They are useful cosmological tools because

their number-density distribution as a function of distance (and hence of cosmic time) sensitively depends on cosmological models, and because the relative mass fraction of their gas and dark matter can be used to determine the total matter density of the Universe (S. D. M. White, MPI für Astrophysik, Garching).

Clusters also distort the shapes and brightnesses of any sources behind them, most spectacularly into the 'giant luminous arcs', highly distorted and magnified images of background galaxies discovered 10 years ago^{3–5}. New observations of the cluster MS1358 with the Hubble Space Telescope (HST) show a strong, very red arc (M. Franx, Kapteyn Inst., Groningen). Spectroscopy using the Keck Telescope implies that it is the lensed image of a galaxy at redshift 4.92. This would not just make it the highest-redshift

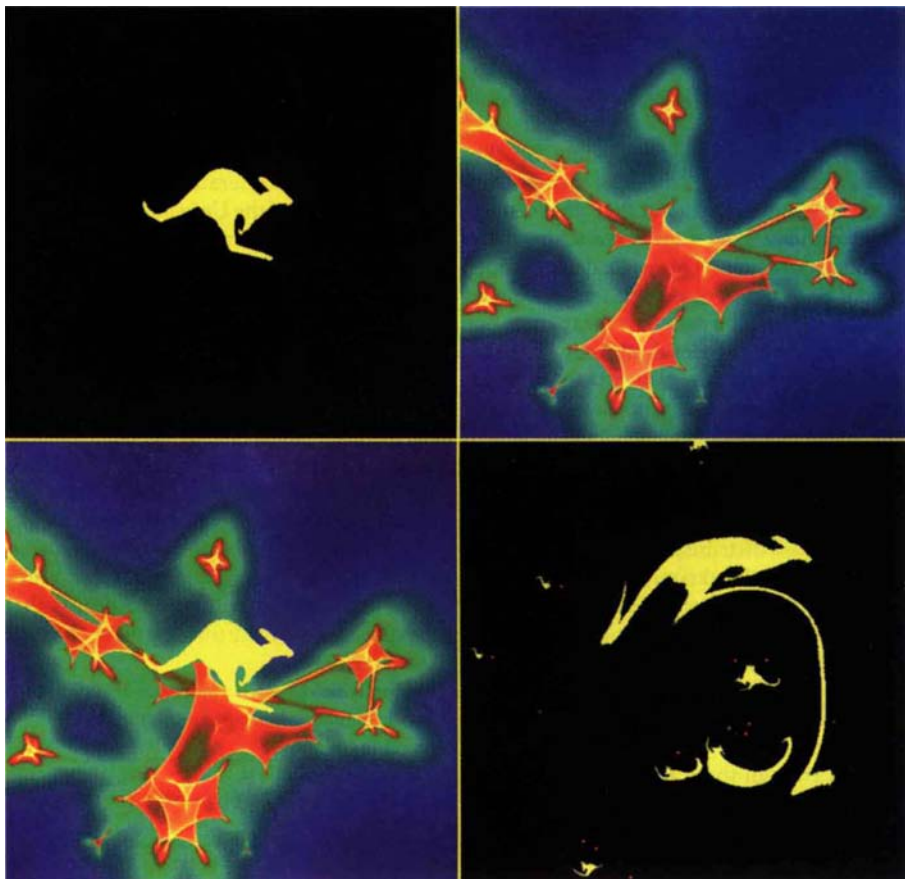


Figure 1 A kangaroo (a) illustrates strong and weak gravitational lensing. About a dozen simulated gravitational lenses produce 'caustic' structures (b), which define regions in the source plane that give rise to multiple images. The yellow lines indicate source positions that correspond to very high magnification, with other colours (red, green, blue) representing increasingly lower magnification. c, The position of the source relative to the caustics. d, The source as seen through the point lenses (marked by red dots). The magnified main image includes a strong giant arc, and there are three other strongly distorted images of the source, and a few fainter mirror-images. Note that the body of the kangaroo is enlarged and distorted, whereas the heel (inside the caustic region) is deformed into a triple arc.

giant arc, but the most distant cosmic object with measured redshift. Two similar red arcs, with redshifts of about 4.05, are seen in the cluster A2390 (R. Pello and G. Soucail, Obs. Midi-Pyrénées, Toulouse).

The mass distribution in the central part of galaxy clusters is still a matter of debate. Most simulations predict that clusters have sharply peaked central density, rather than a finite core where the mass distribution flattens. Fortunately, so-called radial arcs — linear images of background galaxies oriented in the direction to the cluster centre, like a spoke in a wheel — can precisely mark the size of such a core. A second radial arc found in the cluster MS2137 settles its core size at around 30,000 parsecs.

A striking feature of strong gravitational lenses is the formation of multiple images of a background source. In the cluster AC114, two five-image configurations have been discovered (J.-P. Kneib, Obs. Midi-Pyrénées), but even more impressive is the first septuple arc: seven images of the same background galaxy, lensed by the cluster CL2244. This abundance of high-multiplicity images tells us that the matter distribution in clusters is not smooth, but rather that the individual elliptical galaxies have massive dark haloes and hence produce complicated caustics (Fig. 1) and multi-image configurations (Kneib).

Giant arcs are produced only by the biggest galaxy clusters. Smaller clusters are much more numerous, but they only distort background galaxies slightly, stretching their images tangentially. This weak deformation cannot be detected in a single lensed object, but because this is a coherent shear effect on all background galaxies, hundreds or thousands of galaxies are tangentially aligned around the cluster. This weak-lensing signature contains a huge amount of information about the cluster potential⁶. Elegant reconstruction techniques⁷ have now permitted quantitative determination of the shear and the total mass distribution in more than ten clusters (H. Hoekstra, Kapteyn Inst., Groningen; N. Kaiser, CITA, Toronto; Y. Mellier, Institut d'Astrophysique de Paris, Paris; P. Schneider, MPI für Astrophysik; G. Squires, Univ. Berkeley; L. van Waerbeke, Obs. Midi-Pyrénées). Measurement of shear at the one-per-cent level now seems possible — good enough to be another tool for distinguishing between cosmological models.

The University of Hawaii has built an 8,192-by-8,192-pixel CCD camera that made it possible to map three clusters with comparable redshift within one field. Mass reconstruction and subsequent comparison with each cluster's light distribution makes it clear that the mass-to-light ratios of clusters vary by at least a factor of two. A fourth cluster was discovered just by the statistical analysis of weak distortions — the first 'lens-selected' galaxy cluster (G. Luppino, Univ. Hawaii).

Another tool for determining the masses of galaxy clusters is X-ray observation, which only agrees to within a factor of two with the lens-based method. One reason for this apparent conflict may be the fact that clusters selected because of their strong lensing effect are preferentially dynamically active, with lots of substructure, whereas the X-ray based methods are most reliable for clusters with little substructure (M. Bartelmann, MPI für Astrophysik; H. Böhringer and S. Schindler, MPI für Extraterrestr. Physik, Garching).

A more recent application of gravitational lensing is the study of the effect of large-scale structure on background sources. Here, weak lensing comes in two guises: as well as distorting the shapes of extended sources, it affects the apparent brightness of unresolved sources. Inspired by Richard Feynman, Jim Gunn worked out the consequences of both effects 30 years ago⁸. The first effect will, in the next few years, allow us to determine the underlying primordial power spectrum of the matter distribution through measurements of this 'coherent shear' on large angular scales, and hence learn about the initial conditions of the early Universe (A. Stebbins, Fermilab; B. Jain, MPI für Astrophysik).

But not all aspects of lensing are considered a blessing by astronomers. Large-scale structures modify the apparent brightness of distant 'standard candles', which may make it more difficult than originally thought to determine the deceleration parameter (and so the density) of the Universe by measuring the apparent brightness of type-Ia supernovae as a function of redshift. Fortunately, it turns out that for most cosmological models this lens-induced dispersion is smaller than other observational uncertainties (J. Frieman, Fermilab; J. Wambsganss).

The workshop showed that the prospects for the future of lensing are good. Weak and strong lensing around clusters of galaxies and elsewhere will be one of the most useful tools in the next decade for the study of the matter distribution in the Universe. And it may be that in about a dozen years, most astronomers will have to deal with a gravitational-lens-distorted view of the Universe⁹. □

Joachim Wambsganss is at the Astrophysikalisches Institut Potsdam, An der Sternwarte 16, 14482 Potsdam, Germany.

1. Zwicky, F. *Phys. Rev. Lett.* **51**, 290 & 679 (1937).
2. Walsh, D., Carswell, R. F. & Weymann, R. J. *Nature* **279**, 381–384 (1979).
3. Lynds, R. & Petrosian, V. *Bull. Am. Astr. Soc.* **18**, 1014 (1986).
4. Soucail, G. *et al. Astron. Astrophys.* **172**, L14–L16 (1987).
5. Paczynski, B. *Nature* **325**, 572 (1987).
6. Tyson, J. A., Valdes, F. & Wenk, R. A. *Astrophys. J.* **349**, L1–L4 (1990).
7. Kaiser, N. & Squires, G. *Astrophys. J.* **404**, 441–450 (1993).
8. Gunn, J. E. *Astrophys. J.* **147**, 61–72 (1967); **150**, 737–753 (1967).
9. Press, W. H. in *Astrophysical Applications of Gravitational Lensing* (eds Kochanek, C. S. & Hewitt, J. N.) 407–414 (Kluwer, Dordrecht, 1996).

*Max-Planck-Institut für Astrophysik Workshop on Weak and Cluster Gravitational Lensing, Ringberg Castle, Munich, Germany, 19–25 January 1997.

Daedalus

Thinking makes it so!

Are electromagnetic fields a health hazard, and if so, why? One suggestion is ion cyclotron resonance. This effect is best known in plasmas, whose free ions can be excited to orbit around magnetic field lines by radio waves of the same frequency as the ionic orbits. By analogy, ions in the body might orbit around geomagnetic lines of force, if excited by a suitable a.c. field — in this case in the low audio range. Computer screens, power lines and so on emit in this range; and internal orbiting ions might well be bad for you.

This theory, says Daedalus, suggests a splendid new way of exploring the brain and nervous system. A nerve conducts by the sudden collapse of the voltage across its membrane, as sodium ions flood in from outside. Potassium ions are then promptly expelled again, restoring the resting potential ready for the next pulse. The ions traversing the nerve membrane will be constrained by its narrow channels; but the ones displaced on either side will be freely deflected by the magnetic field. At the fastest nerve-firing rate, the ions would be looping round almost continuous orbits at about 500 Hz. So they could absorb a 500 Hz a.c. signal by ion cyclotron resonance. Potassium ions are heavier than sodium ions, so the oscillation will not be exactly sinusoidal. Its absorption band will have a characteristic shape or overtone structure.

DREADCO technicians are now building a special headset to study the brain by ion cyclotron resonance. Its magnetic field gradient will be chosen so that only a specific slice of the brain can resonate at any given applied frequency, and its a.c. electrodes will be shaped to give a strong field only in a limited area of this slice. Accordingly, it will probe the state of activity in one small region of the brain. This region will be scanned round the brain by varying the field and frequency.

Ion encephalography will be far more informative than the old EEG. Not only can it probe very small regions; it can react as fast as the nerve cells themselves. Under computer guidance, it will soon be able to recognize the signatures of thousands of different brain states. The computer will then be accessible to purely mental control. The wearer of the headset will merely have to think of a command; the headset will detect it; the computer will recognize it and carry it out. At last that pestilent invention, the keyboard, will be vanquished. Lower back pain and repetitive strain injury will vanish from our lives; we will command our computers by thought alone.

David Jones

Natural and artificial 'singing' sands

Two types of sand that emit audible sounds on being sheared, by wind or other mechanical means, have been reported throughout the world over the past century¹⁻⁵. They are commonly known as squeaking, or singing, and booming sands. The exact mechanism by which sounds from these sands are produced is still unknown. Natural musical sands are thought to have a silica gel-type surface layer on the sand grains. We show here that this layer is crucial for the singing behaviour, as a commercially manufactured silica gel, an artificial sand material, 'sings'.

Many singing sands are well sorted materials with sizes in the range 100–500 μm , roughly rounded shapes, and a high content of quartz particles^{3,6}, and yet booming sand materials are also found in Kauai, Hawaii, where sands are mostly calcareous in content. We have used Fourier transform infrared (FTIR) spectroscopy to characterize the surface of musical sands^{7,8}. One significant feature of the infrared spectrum was a broad band at $3,400\text{ cm}^{-1}$, extending from $3,700\text{ cm}^{-1}$ to $2,800\text{ cm}^{-1}$. This band may be due to clusters of water in an amorphous silica layer on the surface of the sand grains⁹.

We have previously shown⁸ that sand materials that do not sing can be made to do so by grinding the grains in a mill, periodically removing the fines and renewing the water, leaving a well sorted and surface-polished material. After sufficient grinding, these can achieve the singing or booming state, and they exhibit the characteristic $3,400\text{ cm}^{-1}$ band.

We have now completed similar studies on the singing sands of Kauai and they too show a $3,400\text{ cm}^{-1}$ infrared band, indicating that a similar water-silica layer may exist on the surface of these non-quartz sand materials. As the solubility of silica in water is about 1 mM (ref. 10), the formation of such a siliceous layer even on non-quartz sand grains is possible. This similarity in structure between singing sands of vastly different bulk composition led us to hypothesize that this silica gel-type layer may be part of the mechanism of sound generation in these materials. Indeed, the repeated suggestion in the literature that a certain degree of dryness (and thus a specific water content) of these materials is required for such sands to sing fits nicely into the silica gel-type mechanism and the known deliquescent properties of silica gels.

If this layer on the sand grains was the cause of the singing phenomenon, then commercially available silica-gel materials should also 'sing'. We tested this idea using commercial grade silica gel (Fisher Scientific) with a size range of 200–500 μm . On sharp shaking of a 200-g sample of this material in a glass bottle, a loud sound was emitted, similar to that of natural singing sands. Furthermore, this material gives the characteristic $3,400\text{ cm}^{-1}$ infrared band and an audio beat pattern in the range of 450 Hz, typical of natural singing sands.

These observations lead us to speculate that, as silica gel is a weakly bonding material (highly hydrated commercial silica gel particles adhere), then silica gel-type layers

found in natural sands might allow their particles to stick weakly together. The emission of a coherent wave sound pattern requires a minimal number of such particles to stick together, and such a loose cohesive mass of appropriately sized sand particles moving through the air, either tumbling down a dune or being shaken in a container, could then act as a resonator.

The finding that commercially available silica gel is a 'musical sand' should enable a more detailed study of the phenomenon, and could lead to practical application of these readily available materials in energy converter devices.

Douglas E. Goldsack

Marcel F. Leach

Cindi Kilkenny

Centre in Mining and Mineral Exploration Research, and Department of Physics, Laurentian University, Sudbury, Ontario P3E 2C6, Canada

1. Carus-Wilson, C. *Nature* **44**, 322–323 (1891).
2. Thomas, E. R. & Jones, J. M. *Victorian Nat.* **81**, 218–224 (1964).
3. Brown, A. E., Campbell, W. A., Jones, J. M. & Thomas, E. R. *Proc. Univ. Newcastle-upon-Tyne Phil. Soc.* **1** (1), 1–21 (1964).
4. Campbell, W. A., Jones, J. M. & Thomas, E. R. *Proc. Univ. Newcastle-upon-Tyne Phil. Soc.* **1** (14), 170–177 (1968).
5. Lindsay, J. E., Criswell, D. R., Criswell, T. L. & Criswell, B. S. *Geol. Soc. Am. Bull.* **87**, 463–473 (1976).
6. Leach, M. F. & Rubin, G. J. *Acoust. Emiss.* **11**, 19–20 (1993).
7. Leach, M. F., Goldsack, D. E. & Chartrand, H. J. *Proc. 8th Pacif. Conf. NDT* 55–62 (Taipei, Taiwan, 1995).
8. Leach, M. F., Goldsack, D. E. & Kilkenny, C. *Proc. 6th Conf. Acoust. Microseismic Activ. Geol. Struct.* (Pennsylvania State Univ. Press, in the press).
9. Kronenberg, A. in *Reviews in Mineralogy* Vol. 29 (eds Meaney, P. H., Prewitt, C. T. & Gibbs, G. V.) Ch. 4 (Mineralogy Soc. Am., Washington DC, 1994).
10. Henderson, P. *Inorganic Geochemistry* 256–257 (Pergamon, New York, 1982).

Human haemoglobin from transgenic tobacco

Haemoglobin-based blood substitutes are currently being developed^{1,2} using haemoglobin from sources including outdated human or bovine blood, and expression in bacteria, yeasts or transgenic animals³. The oxygen affinity and stabilization of the purified haemoglobins can be controlled by a combination of crosslinking and site-directed mutagenesis^{3,4}, but other problems, such as haem oxidation and the presence of infectious agents, remain to be solved. Transgenic plants may be a suitable alternative source of haemoglobin as they constitute an inexpensive source of biomass, and transgenic crops are becoming commercially available⁵. Here we demonstrate the potential of this approach by co-expressing the α - and β -globins of the human haemo-

globin HbA in transgenic tobacco plants, leading to the production of functional tetrameric haemoglobin.

We fused coding sequences of α - and β -globins to the sequence of the chloroplast transit peptide (containing the first 57 residues of the protein precursor) of the small subunit of Rubisco from *Pisum sativum* L. (ref. 6). Co-expression was obtained with the binary plasmid pBIOC59 expressing fused α and β sequences. This

plasmid is derived from pGA492 (ref. 7) into which two expression cassettes, 'Pd35S-T35S' derived from pJIT60 (ref. 8), were inserted. Tobacco plants (*Nicotiana tabacum* var. Xanthi) were transformed using *Agrobacterium tumefaciens* strain LB4404 (ref. 9). We screened different plant-organ extracts from regenerated transformants for the presence of α - and β -globin chains by western blotting¹⁰.

Globin chains were detected in extracts

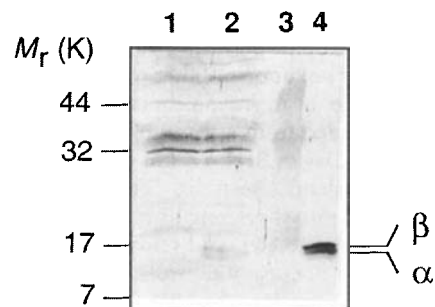


Figure 1 Western blot detection of the α - and β -globins in crude seed extracts from transgenic tobacco plants. Lane 1, 75 μg wild-type extracted plant protein. Lane 2, 75 μg transgenic extracted plant protein. Lane 3, molecular mass markers. Lane 4, 50 ng purified human HbA. The western blot analysis was done with rabbit anti-human haemoglobin immunoserum and goat anti-rabbit IgG antibody coupled to alkaline phosphatase.

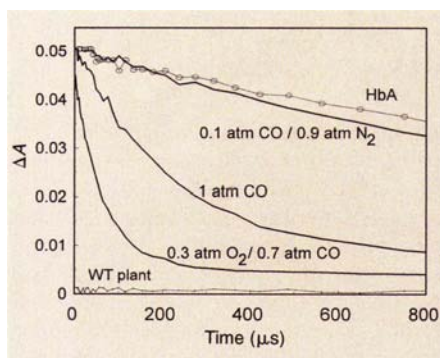


Figure 2 Ligand recombination kinetics to rHb samples extracted from transgenic plants, equilibrated under 0.1 or 1 atm CO, or a mixed CO/O₂ atmosphere. Also shown are the kinetics for HbA added to plant extracts and results from control wild-type (WT) plants (0.1 atm CO). ΔA is the change in absorbance. Photolysis used 10 ns pulses at 532 nm; detection at 436 nm.

from seeds (up to 0.05% total extracted proteins) and roots in over half of the plants that were transformed (Fig. 1). Recombinant globins have a similar molecular mass to the native globins, indicating that the transit peptide was properly cleaved. We extracted recombinant haemoglobin (rHb) from tobacco seeds and purified it by two successive chromatography steps.

The functional properties of the rHb were studied by flash photolysis, as the kinetics of ligand recombination with haemoglobin tetramers display a characteristic biphasic shape. HbA purified from human blood added to extracts from wild-type seeds showed typical biphasic kinetics for carbon monoxide rebinding (Fig. 2), showing that haemoglobin was stable in this environment. Recombinant haemoglobin purified from seed extracts had similar biphasic recombination kinetics to those of HbA (Fig. 2). A re-equilibration of the sample under 1 atm CO showed the expected acceleration of bimolecular kinetics (Fig. 2). As for HbA, oxyhaemoglobin samples provided only weak transient absorption signals, but use of a mixed CO/O₂ atmosphere allowed photodissociation of the stable HbCO form, followed by rapid oxygen binding (Fig. 2). A very slow phase, taking 1 s (not shown), corresponds to the replacement of O₂ with CO.

These results show that functional recombinant human haemoglobin, a complex multimeric protein, can be obtained from plant sources. Although tobacco plants were used as a model system, other species may produce higher yields with simpler purification and sterilization steps. Transgenic plants offer considerable advantages for the production of large quantities of rHb and would alleviate the dependence on limited supplies of outdated human

blood as well as eliminating contamination from bacterial and animal sources.

Wilfrid Dieryck

Josée Pagnier

Claude Poyart

Michael C. Marden*

Inserm U299, Hôpital de Bicêtre,
94275 Le Kremlin-Bicêtre, France
e-mail: marden@kb.inserm.fr

Véronique Gruber

Philippe Bournat

Sylvie Baudino

Bertrand Mérot

Biocem, Groupe Limagrain,
Campus Universitaire des Cèzeaux,
63170 Aubière, France

1. Winslow, R. M. *Hemoglobin Based Red Cell Substitutes* (Johns Hopkins Univ. Press, Baltimore, 1992).
2. Kumar, R. *et al. Proc. Soc. Exp. Biol. Med.* **208**, 150–158 (1995).
3. Looker, D. *et al. Nature* **356**, 258–260 (1992).
4. Baudin, V. *et al. Biophys. Biochem. Acta* **1159**, 223–226 (1992).
5. Grierson, D. *et al. in Genetic Engineering of Crop Plants* (eds Lycett, G. W. & Grierson, D.) 115–125 (Butterworths, London, 1990).
6. Anderson, S. & Smith, S. *Biochem. J.* **240**, 709–715 (1986).
7. An, G. *Plant Physiol.* **81**, 86–91 (1986).
8. Guerinneau, F. & Mullineaux, P. in *Plant Molecular Biology Labfax* (ed. Croy, R. R. D.) 121–147 (BIOS Scientific, Oxford, 1993).
9. Horsch, R. B. *et al. Science* **227**, 1229–1231 (1985).
10. Sambrook, J., Fritsch, E. F. & Maniatis, T. in *Molecular Cloning* (ed. Nolan, C.) (Cold Spring Harbor Laboratory Press, New York, 1989).

* To whom correspondence should be addressed.

Laser measurements of coral growth

Calcification in reef-building corals is highly variable and strongly affected by environmental factors¹, especially light, temperature, water motion and pollution². The mechanisms by which environmental influences alter calcification are not well understood^{3–5}, but it is widely accepted that calcification is somehow enhanced by photosynthesis of algal symbionts, known as zooxanthellae, living within the cells of reef-building corals⁶. We describe here the use of a laser, underwater, to measure skeletal extension (the process by which the coral skeleton increases in size, rather than calcification, which includes both extension and thickening of skeletal elements^{2,3}) in the Great Barrier Reef staghorn coral, *Acropora grandis* Brook, over a timescale of minutes to hours. Our method is non-destructive and allows us to return to the same colony over and over again, in different seasons, after catastrophic events, or at the onset or removal of pollution or other stressors.

Branches of *A. grandis* about 15 cm long were collected from a depth of 6 m in the lagoon at Davies Reef, 77 km off the coast of Queensland, Australia. The branches were fixed into short lengths plastic tubing, 12 mm in diameter, and remained undisturbed for at least four weeks before the measurements. We then measured linear

extension of the branch tip from changes in the distances between lines in the diffraction pattern formed by a laser beam on a screen behind the coral.

A 3 mW semiconductor laser diode (No. 194-026; RS components, Australia) was contained in an underwater housing and produced a ~3 mm-diameter beam of 670 nm red light. The laser was supported by a hydraulic positioning mechanism to allow fine adjustments to the beam alignment (Fig. 1). The beam was aimed through a 0.3–1.0 mm 'slit' created between the growing tip of the branch and a 'reference-plate' positioned above the tip only during the measurement. The distance between the diffraction lines increased as the tip grew towards the reference plate.

These distances were converted to growth distances from a linear calibration using slits of known width, also obtained *in situ*, allowing us to resolve growth increments as small as 10 μ m. Tips were exposed to laser light for 30–60 s (ref. 7), and the reference plate was withdrawn thereafter. After between 2 and 12 h, the reference plate was lowered back to its exact initial position, and a second measurement was made. Throughout the measurement period, irradiance and temperature were

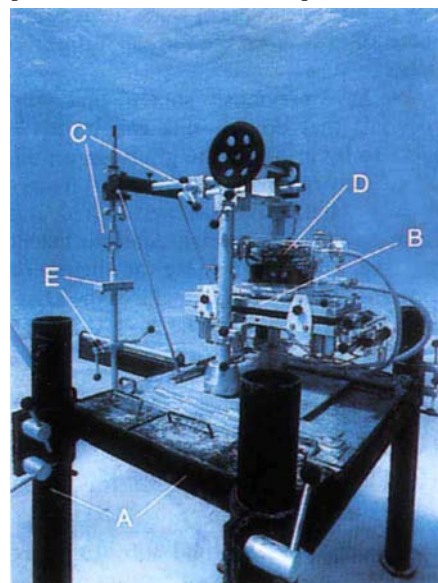


Figure 1 Our laser apparatus is pictured measuring growth in the staghorn coral, *Acropora grandis* on the Great Barrier Reef. A, Heavy anchoring and supporting frame. B, System for aligning the laser beam — the laser is initially positioned manually and then fine adjustments in 3 axes are made using knobs that operate hydraulic pistons. C, Lowering mechanism, which forms the 'slit' between the reference-plate and the tip of the coral branch. A diffraction pattern is projected onto a white screen (not shown) when the beam passes through the slit. D, Control module containing the laser and data logger, which records temperature and light. E, Coral positioning pedestal.

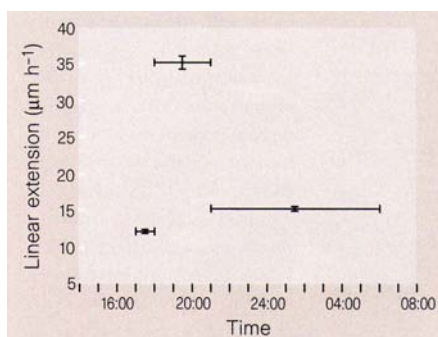


Figure 2 Diel skeletal extension pattern of *A. grandis* at the Great Barrier Reef. No growth occurred in the period that is not represented in the graph. The horizontal bars represent the time period measured, $n=6$ and the error bars are shown. The laser apparatus was positioned on top of an axial corallite and measurements were taken underwater at 16:00, 18:00, 21:00 and 6:00 hrs.

constantly recorded.

The measurements described here were taken between 5 and 11 May 1996, when water temperature was 28 °C. We found large diel differences in growth rate (Fig. 2). Except at dusk and dawn, daytime measurements showed no extension. Extension took place at night, when rates averaged $15 \pm 0.36 \mu\text{m h}^{-1}$ ($n=6$), stopping soon after sunrise and resuming just before dusk. Peak values as high as $35 \pm 1.75 \mu\text{m h}^{-1}$ ($n=6$) were reached roughly 30 min before sunset and lasted for 2 or 3 h. Using the same technique we found a similar diel pattern of extension in a massive species, *Porites lutea* Edwards and Haime.

Our findings agree with those of ref. 8, which suggests that skeletal extension is not dependent upon concomitant daytime photosynthesis of the zooxanthellae. We suggest that skeletal accretion in corals and extension are not necessarily coupled. This fits well with recent models of growth in long-lived, massive coral colonies⁹. These results, and the finding that extension occurs chiefly during the hours after sunset, have considerable implications for determining climate and environmental information from corals.

R. Vago, E. Gill

J. C. Collingwood

Australian Institute of Marine Science,
PMB No. 3, Townsville,
Queensland 4810, Australia

DNA repair by recycling reverse transcripts

Teng *et al.*¹ and Moore and Haber² recently reported that in the yeast *Saccharomyces cerevisiae*, double-stranded chromosomal breaks were often repaired by the insertion of complementary DNA, generated by reverse transcription from RNA sequences by retroelements expressed in those cells. This process was identified by analysing the HO endonuclease-recognition site in the mating-type (*MAT*) locus of cells that survived expression of the HO nuclease, and occurs mainly by non-homologous recombination. Here I propose an alternative, and to my mind more plausible, explanation for some of the results presented.

The sequences that were inserted in several of the reported repaired breaks originated from the yeast retrotransposon Ty1 (refs 1, 2). The sequences frequently spanned a part of the Ty1 long-terminal repeat (LTR) element, or the whole 3' LTR and the upstream Ty1 reporter *HIS3* gene³, and a part of the primer-binding-site sequence. The primer-binding-site region is complementary to the initiator methionine transfer RNA used as the primer for Ty1 minus-strand DNA synthesis (Fig. 1). Both reports suggested that the primer tRNA moiety was directly involved in the recombination events, because these structures resemble the minus-strand strong-stop DNA (a single-stranded DNA of the R and U5 region, with primer tRNA still attached, see Fig. 1) or the minus-strand DNA after strand transfer (in experiments using Ty1 containing the *HIS3* marker gene).

I find such an explanation unlikely, because full-length minus-strand strong-stop DNA is not seen *in vivo*³ and because the attached primer tRNA is likely to be folded into a very stable tertiary conformation and so would not be accessible to the recombination machinery. I propose that the tRNA sequence is instead provided by the Ty1 plus-strand DNA either in a single-stranded form or in a duplex with the minus strand. The Ty1 reverse transcriptase copies part of the primer tRNA sequence very efficiently into the plus-strand DNA during plus-strand strong-stop formation⁴ — only this part of the tRNA sequence has been found in the HO site. This strong-stop DNA is the most abundant product of the Ty1 reverse transcription process^{4,5}, but is not used in further Ty1 synthesis⁶. I suggest that yeast can 'recycle' these unused cDNAs in processes such as the repair of double-stranded chromosomal breaks.

I make two predictions from this hypothesis. First, the frequency of the *HIS3* insertion events should increase on deletion of the

Ty1 PPT1 sequence (Fig. 1f), as this region gives rise to a primer for plus-strand synthesis. The *HIS3* gene is inserted in Ty1 upstream of the PPT1 sequence, therefore when PPT1 is present most repairs would use DNA that spans only the LTR and not *HIS3*. Second, the insertions should not contain tRNA sequences when both PPT1 and PPT2 are deleted (Fig. 1g; PPT2 is the second primer for plus-strand synthesis, located around the middle of the Ty1 genome), as this would abolish synthesis of plus-strand DNA^{4,5}.

If tRNA is involved in the double-stranded break repair process *in vivo*, this would indeed have important biological implications⁷, but this is not yet proven. Therefore, I offer an alternative, perhaps simpler, explanation for the results of refs 1 and 2, namely that during the formation of Ty1 plus-strand DNA, reverse transcription

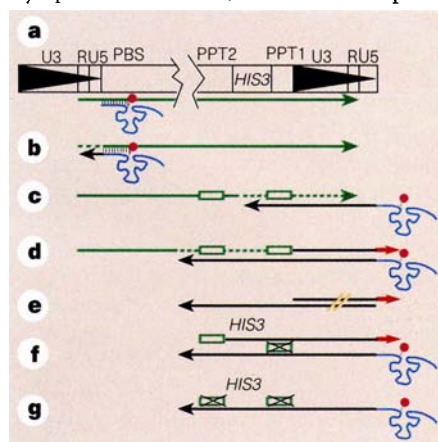


Figure 1 Schematic diagram of Ty1 reverse transcription. a, Structure of Ty1 DNA (black) and its RNA transcript (green) initiating and ending within long-terminal repeats. Primer initiator methionine tRNA (blue) is annealed to the primer-binding site (PBS). b, The minus-strand strong-stop is completed when DNA synthesis reaches the end of the RNA template. This RNA bound to the newly synthesized DNA is degraded by RNaseH (dashed line). c, First (minus) strand transfer is facilitated by the repeated sequence, R, found in the LTR. The minus strand is then extended along the RNA template. d, The primer for plus-strand synthesis is derived from the purine-rich sequence PPT1 (rectangle). Synthesis proceeds into the tRNA sequence (red arrow). The stop signal for reverse transcription is the first modified base to be encountered (indicated in purple). The tRNA, now in a duplex with plus-strand DNA, is removed by RNaseH. e, Proposed intermediate used for the double-strand break repair. Yellow lines represent a possible break or recombination at a microhomology in Ty1 DNA to the recipient ends. f, Deletion of PPT1 would increase the efficiency of *HIS3* insertion into the HO cleavage site. g, Deletion of PPT1 and PPT2 (which encodes a second primer for plus-strand DNA synthesis) should almost eliminate *HIS3* insertions containing the tRNA sequences by decreasing Ty1 plus-strand synthesis.

1. Barnes, D. J. & Chalker, B. E. in *Coral Reef Ecosystems* Vol. 25 (ed. Dubinsky, Z.) 109–131 (Elsevier, Amsterdam, 1990).
2. Barnes, D. J. & Lough, J. M. *J. Exp. Mar. Biol. Ecol.* **126**, 119–134 (1989).
3. Gladfelter, E. H. *Proc. 5th Int. Coral Reef Congr.* 527–539 (1985).
4. Buddemeier, R. W. & Kinzie, R. A. *Oceanogr. Mar. Biol. Ann. Rev.* **14**, 183–225 (1976).
5. Marshall, A. T. *Science* **271**, 637–639 (1996).
6. Goreau, T. F. & Goreau, N. I. *Biol. Bull.* **117**, 239–250 (1959).
7. Stromgren, T. *Limnol. Oceanogr.* **20**, 845–848 (1975).
8. Barnes, D. J. & Crossland, C. J. *Limnol. Oceanogr.* **25**, 1113–1117 (1980).
9. Taylor, R. B., Barnes, D. J. & Lough, J. M. *J. Exp. Mar. Biol. Ecol.* **167**, 109–125 (1993).

of part of the tRNA occurs, providing the genetic material that is inserted by non-homologous recombination events.

Vit Lauermaann

Advanced BioScience Laboratories,
Frederick Cancer Research and Development Center,
Building 539, PO Box B,
Frederick, Maryland 21702, USA
e-mail: vit@welchlink.welch.jhu.edu

Haber and Moore reply—We do not yet know how a specific segment of Ty1 can be captured at the sites of double-strand breaks, and the suggestion by Lauermaann provides one plausible explanation of the source of the putative cDNA segment. But we feel that the details of the events that we characterized favour our original interpretation.

The intermediate suggested by Lauermaann would contain cDNA initiated at the primers PPT1 or PPT2 and should contain all of the long-terminal repeat (U3-R-U5), but three of the four repairs we studied contained R and U5 but lacked any U3. We prefer an explanation in which the initial minus-strand strong-stop cDNA, beginning at the primer-binding site and including only U5 and R, was the source of the captured DNA. The abundance of this intermediate might be greater among (often mutant) natural Ty1 elements than for a construct carrying an inducible Ty1 gene. This is consistent with our observation that expression of Ty1 messenger RNA from a galactose-inducible promoter had little effect on the frequency of these largely endogenous events².

Lauermaann's explanation is more likely for the insertion of *HIS3* derived from cDNA of a galactose-inducible Ty1 *mhis3AI* construct observed by Teng *et al.*¹, but more recent results from Teng and Gabriel suggest that this may not be the case (see below).

The covalent ligation of RNA with DNA during insertion by DNA replication is not entirely new. It has recently been documented for group II introns in a reverse splicing reaction⁸. But such an RNA–DNA junction is not a necessary intermediate in the process. It would suffice for the tRNA–cDNA segment to be used as a template by one 3' end of the broken DNA, to copy these sequences. The fact that all of the Ty1 sequences containing R and U5 are inserted in one orientation supports this idea (ref. 2 and see below).

Clearly additional experiments are needed to determine which cDNA intermediate is used as the source of the inserted DNA, and to establish how the non-homologous junctions are formed.

James E. Haber, J. Kent Moore

Rosenstiel Center MS029, Brandeis University,
Waltham, Massachusetts 02254-9110, USA

Teng and Gabriel reply—Lauermaann has suggested that the observed^{1,2} Ty1 insertions associated with the primer-binding site may be derived from plus-strand cDNAs, whose

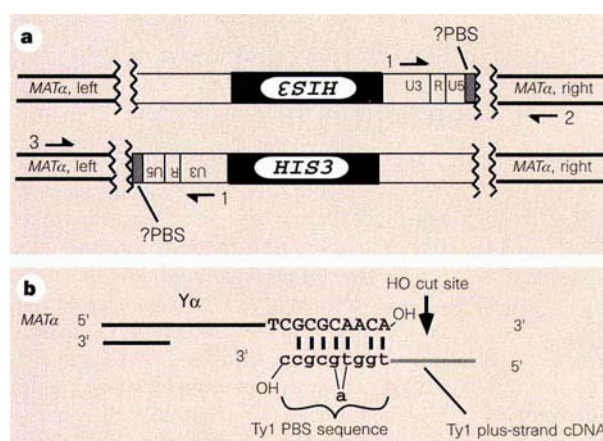


Figure 2 Preferential association of Ty1 primer-binding-site (PBS) sequences with one side of the HO cut site in *MATα*. a, The two possible orientations of *HIS3* insertions at the HO cut site in *MATα*, and a PCR strategy for identifying the orientation of insertions in which the Ty1 U3 sequence is closely linked to the *MAT* locus. b, Potential alignment of the 3' terminal bases to the left of the HO cut site in *MATα* and the Ty1 plus-strand cDNA terminating in the PBS.

3' ends are templated by primer tRNA present at the 5' end of minus-strand cDNA. Further, he suggests specific experiments to verify his hypothesis. Recent data from our laboratory bear directly on these proposals.

We examined histidine prototroph formation among survivors of HO cutting in yeast strains carrying wild-type, *ppt1*[−] mutant, or *ppt1*[−]/*ppt2*[−] double mutant versions of a *GAL1*-Ty1 element marked with *mhis3AI*. We screened 52 independently targeted histidine prototrophs (Fig. 2a) by polymerase chain reaction (PCR) and identified nine (17%) in which U3 sequences were closely linked to either the right (Z1) or left (Yα) side of the HO cut site at *MATα*. A strong bias for linkage with the left side of the HO cut site was seen (eight of nine prototrophs, Table 1). By sequencing these eight junctional segments, we determined that each contained between four and ten bases of the primer-binding site separating U5 from the *MAT* segment. These sequences were indistinguishable from those observed by Moore and Haber² in wild-type yeast. A potential base complementarity of 7/10 between the left end of the HO cut site and the primer-binding site (Fig. 2b) may underlie this strong insertion bias.

Our findings do not support Lauermaann's prediction that insertions derived from the *ppt1*[−]/*ppt2*[−] mutant version of Ty1 will lack associated primer-binding-site sequences. Instead our results suggest that minus-strand cDNA with its attached tRNA—the expected Ty1 replication intermediate in the *ppt1*[−]/*ppt2*[−] double mutant

strain—is the most likely substrate for insertional repair.

A plausible model for repair in these cases would involve plus-strand synthesis initiated from the right end of the cut site, using minus-strand cDNA as the template. Plus-strand synthesis would end with partial copying of the attached tRNA. After RNaseH cleavage of the tRNA–DNA duplex, the terminal primer-binding-site region would be exposed and could be anchored to the left end of the HO cut site by complementary base pairing. Our data indicate that, even with wild-type Ty1 constructs, insertions may be formed from intermediates or 'dead-end' products of the replication process, as ends containing primer-binding sites are not expected to be found in mature double-stranded Ty1 DNA.

Our new results cannot distinguish between repairs initiated from plus-strand or minus-strand cDNA, or even from RNA templates. However, the primer-binding site–HO cut site complementarity suggests that the observed junctional 'hotspot' results from terminal base pairing and consequent stabilization of the two single-strand ends. Yeast may use various substrates during insertional repair of double-strand breaks, and our data show that minus-strand Ty1 cDNA linked to its primer tRNA might be one such intermediate. Further work is needed to determine the relative contributions of different repair substrates, including plus-strand cDNA, as suggested by Lauermaann.

Shu-Chun Teng

Abram Gabriel

Department of Molecular Biology and Biochemistry,
Rutgers University, Piscataway,
New Jersey 08855, USA

Table 1 Results from PCR analysis

<i>GAL1</i> -Ty1 associated element	No. analysed	No. associated with <i>MATα</i> , L	No. associated with <i>MATα</i> , R
Wild type	26	3	1
<i>ppt1</i> [−] mutant	11	1	0
<i>ppt1</i> [−] / <i>ppt2</i> [−] mutant	15	4	0

Tabulation of results from the PCR analysis shown in Fig. 2a, using independent *MATα* targeted insertion events derived from a variety of Ty1 constructs, carrying wild-type, *ppt1*[−] mutant or *ppt1*[−]/*ppt2*[−] mutant versions of a *GAL1*-Ty1 element marked with *mhis3AI*. L, left (primers 1 and 3); R, right (primers 1 and 2). Targeted *HIS3* insertions were obtained as described in ref. 1.

1. Teng, S.-C., Kim, B. & Gabriel, A. *Nature* **383**, 641–644 (1996).
2. Moore, J. K. & Haber, J. E. *Nature* **383**, 644–646 (1996).
3. Pochart, P. *et al.* *Nucleic Acids Res.* **21**, 3513–3520 (1993).
4. Lauermaann, V., Nam, K., Trambly, J. & Boeke, J. D. *J. Virol.* **69**, 7845–7850 (1995).
5. Heyman, T. *et al.* *J. Mol. Biol.* **253**, 291–303 (1995).
6. Lauermaann, V. & Boeke, J. D. *Proc. Natl. Acad. Sci. USA* **91**, 9847–9851 (1994).
7. Derr, L. K., Strathern, J. N. & Garfinkel, D. *J. Cell Biol.* **67**, 355–364 (1991).
8. Yang, J., Zimmerly, S., Perlman, P. S. & Lambowitz, A. M. *Nature* **381**, 332–335 (1996).

Cinemagmatic mayhem

Dante's Peak

A Roger Donaldson film
Universal Studios: 1997

Jonathan Fink

The first signs of an impending volcanic calamity can be hard to recognize. Dead trees and wildlife, cloudy tap water and unusual smells in the forest may all be explained by a variety of natural causes other than the stirrings of a magma chamber. Volcanologists who have to decide if such signals are benign or premonitory can face enormous personal and professional pressure. A false alarm can cost them their credibility and cost the local population millions of dollars in lost business; failure to predict a major eruption can result in tens of thousands of casualties. This dilemma lies at the heart of a spate of new films and television programmes.

As backdrops to melodrama, volcanic events are a scriptwriter's dream. Set in rugged terrain with dazzling scenery, most eruptions start slowly and ambiguously, building to a climax (or returning to calm) over periods of days, weeks, months or years. This uncertainty sets up potentially deadly conflicts between heroes (scientists and park rangers) who predict disasters, and villains (developers and politicians) who counsel a more cautious attitude. Chase-scene possibilities abound, with angry lava, pyroclastic flows, mudflows and landslides nipping at the heels or bumpers of fleeing, defenceless humans. Contrast all this potential with the more instantaneous, one-dimensional threats of earthquakes, asteroid impacts and tornadoes, and it is clear why they are becoming the movie moguls' disaster of choice.

Dante's Peak uses a volcano like Mount St Helens in the northwestern United States as a backdrop for a threatened town; US Geological Survey bosses and a greedy helicopter pilot as villains; a rogue British volcanologist as the dashing, misunderstood hero; and two children, a dog and a grandmother as victims in dire need of rescue. Disagreement among scientists about how to interpret geochemical and earthquake data delays the town's evacuation, guaranteeing a crisis at the film's core. The star, Pierce Brosnan, emulates his earlier James Bond role by outwitting the volcano at every turn while charming (and saving) the town's embattled mayor and her family.

Volcanologists got their first hint of this onrush of simmering celluloid last summer when a flurry of messages on an e-mail bulletin board called *Volcano Listserv* indicated that 1997 could be their profession's warholian 15 minutes of fame. Jack Lockwood, one of three consulting geologists for *Dante's Peak*, claimed it would have unprecedented scientific accuracy, educating its viewers while terrify-

Magma cum laude: Dante's Peak gets high marks from volcanologists.

ing them. Most of the volcanological community remained sceptical, especially when reports started circulating about another major production (*Volcano*, 20th Century Fox, scheduled for US release in May 1997) that has lava gushing out of the La Brea tar pits and spreading across downtown Los Angeles.

It may be that *Dante's Peak* will be enjoyed more by the experts than by the targeted teenage audience. There are dozens of anecdotal references to recent volcanological events, primarily through asides that will sound like idle chatter to most of the audience but will trigger vivid memories for regular volcano-watchers. Similarly, the subtle warnings that help Brosnan to sense an approaching disaster may be dismissed as implausible by people ignorant of volcanic processes, but will send shivers down the spines of those who have faced such situations themselves.

Not that the film is free from inaccuracies, however. The eruption is 'Hawaiified', with torrents of basaltic lava crashing through cabins and blocking escape routes. Although ancient basalts are found around many of the Cascade volcanoes, the simultaneous effusion of fluid magma streams and paroxysmal explosive eruptions has never been documented. Even more incredible is that Brosnan's four-wheel-drive vehicle spends several minutes traversing an active basalt flow without bursting into flames. (One of Lockwood's few complaints about the producers was their willingness to limit this lava-crossing episode to a more believable five seconds.) But even with these faults, the large audience and excellent computer effects should ensure that it enlightens more of the public about volcanic dangers than any previous film.

Dante's Peak, the television film *Fire on the Mountain* (ABC) and the forthcoming *Volcano* might have been inspired by real events. The 1991 mega-eruption of Mount

Pinatubo in the Philippines, which was predicted by a team that included the seismologist David Harlow, one of the consultants on *Dante's Peak*, featured in the 1993 film *In the Path of the Killer Volcano* (Nova).

One of *Dante's Peak's* finest accomplishments, from both technical and educational standpoints, is its simulation of roiling ash-flow clouds racing down the mountain's flanks. Few non-scientists are aware of these scorching blasts, and only a handful of people have survived encounters close enough to allow them to evaluate the accuracy of these special effects. Perhaps the best records of these and practically all other eruptive phenomena were made by the late cinematographers Maurice and Katia Krafft, a daring French couple who perished in 1991 while documenting pyroclastic flows at Mount Unzen in Japan. Although not trained as scientists, the Kraffts have protected more people from volcanic dangers than anyone else. Their films, including *Understanding Volcanic Hazards* and *Reducing Volcanic Risk*, have been shown to threatened villagers and townspeople around many of the world's most notorious volcanoes, and persuaded residents to prepare for possible future evacuation. Yet the Kraffts never took themselves as seriously as this year's 'cinemagmatists'. Maurice used to joke about building an asbestos and titanium kayak to slalom down active lava flows. It's a pity he never had the chance to do this as a cameo role in a Hollywood blockbuster. □

Jonathan Fink is at Arizona State University, Tempe, Arizona 85287, USA. Readers interested in a free subscription to *Volcano Listserv* should use his e-mail address, jon.fink@asu.edu. *Understanding Volcanic Hazards* and *Reducing Volcanic Risk* are available for \$19.95 each from NWIA, 3029 Spirit Lake Highway, Castle Rock, Washington 98611, USA.

The golden age of chemistry

Laws and Order in Eighteenth-Century Chemistry

by Alistair Duncan

Oxford University Press: 1996. Pp. 253.

£50, \$90

A History of Chemistry

by Bernadette Bensaude-Vincent and Isabelle Stengers

Harvard University Press: 1997. Pp. 294.

\$35, £25

David Knight

Those who browse among the chemistry books of the late eighteenth century may have noticed that they begin with an account of the attractions of aggregation and affinity between particles that has no connection with the subsequent experiments. Alistair Duncan is concerned with this question, seeing eighteenth-century chemists as being preoccupied with achieving respectability as men of science rather than being seen as artisans, or perhaps charlatans. They therefore borrowed a theory of matter from aristotelian or newtonian physics while building an experimental science that went beyond mere empiricism.

Duncan explores the shifting frontier between chemistry and physics in a series of thematic chapters. The core of the book is a close examination of the tables of affinity and associated diagrams with which chemists hoped to quantify and predict reactions. But no matter how they were extended and modified, such tables were of little use in the laboratory — it was through weighing and the manipulation of gases that chemistry became a quantitative science independent of physics. Its autonomy was asserted by Antoine Lavoisier, who in 1789 dismissed atoms or particles as being metaphysical.

Duncan's is a valuable study, based on wide and close reading of primary sources, and engages seriously the perceptions of his protagonists. Though written for those with some knowledge of the science and the time, it is nonetheless lucid, and will interest and inform anyone who wants to get away from too simple an account of the 'goodies' and 'baddies' in Lavoisier's chemical revolution. It may be old-fashioned in its attention to texts, in that Duncan has chosen to study the writings of chemists, but he is also clearly aware of their context.

In recent years many histories of chemistry have been written, with authors taking refreshingly different perspectives. Bernadette Bensaude-Vincent and Isabelle Stengers are professional historians, and have written a book that will be accessible

to those with little chemistry, and stimulating to those better informed. They reject the commonplace idea that there is some unchanging essential thing called chemistry, about which researchers are finding out more and more. From the beginning, they see chemistry's place on the intellectual map as determined by constantly changing laboratory techniques, professions and institutions. What could those who called themselves chemists make or measure? What jobs did they do? How were they organized? This pattern is still shifting. In many ways the golden age of chemistry, its period of autonomy, was the 'long nineteenth century' (1789–1914), but the authors believe that its history is still being made in terms of these three factors and their interaction.

There are therefore no details of the Hofmann degradation reaction or of Kipps' apparatus, and there are almost no equations or formulae. Nevertheless, by skilled focusing on particular episodes and by appropriate international comparisons, the authors do not make us feel rushed as we move from the heyday of alchemy to the late twentieth century. It is, after all, an academic virtue to know what to leave out. We are given some of the hard science, and we meet chemists making their careers, but most of all there is a historical analysis of chemistry as something that has been transformed over its long history.

Here too we find scientists searching for respectability, at a time when alchemy and something like our chemistry could not be distinguished, and Robert Boyle's scepticism about earlier chemical theories seemed appropriate.

A chapter called "Revolution!" takes us through an eighteenth century much like Duncan's to Lavoisier. The authors take the theory of phlogiston seriously, and regard the synthesis of water from Lavoisier's 'oxygen' and 'hydrogen' as most decisive in overthrowing it. However, they point out that "no chemical theory has imposed itself on its own ... behind each truth we discover campaigns to publicize and promote it, networks of alliances, and processes of selection"; these are duly described.

Lavoisier the taxman/chemist is seen as ending an era. The chapters on the nineteenth century look at two aspects: 'a science of professors' and 'industrial expansion'. The first of these shows that chemists achieved not merely respectability but dignity. The second, under the headings 'analysis', 'substitution' and 'synthesis', describes the practices, goals and theories that guided the chemists of the nineteenth century. These chapters also indicate the importance of networks, societies and journals; of generation

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Lavoisier: asserted the autonomy of chemistry.

gaps with professors supporting, or not supporting, their brightest students; and the creative and strategic implications of synthetic chemistry.

Paradoxically, the very usefulness of chemistry gave a boost to the pure science — the distinction between pure and applied chemistry would have meant nothing to Humphry Davy, for example — as chemical industry went from the Leblanc process for making sodium carbonate through agriculture and explosives to the manufacture of aluminium, rubber and plastics. Justus Liebig and August Hofmann were happy to see their pupils enter industry, such was its importance.

The story ends with the chemistry of the twentieth century 'reduced' to being just a kind of applied physics. Although still important, it has lost the fundamental character it seemed to have in the early nineteenth century, when hundreds flocked to lectures by Davy and his contemporaries. But having lost an empire, chemistry has found a role as the service science everybody needs, in biochemistry, biotechnology and synthesis. Bensaude-Vincent and Stengers see hope of a new territory in Ilya Prigogine's thermodynamics of far-from-equilibrium systems, but whether they are good prophets only time will tell. We can say that theirs was a good story well told — like all good history, it is controversial and makes the reader think about past and present. □

David Knight is in the Department of Philosophy, University of Durham, 50 Old Elvet, Durham DH1 3HN, UK.

Elementary cycling

Cycles of Life: Civilization and the Biosphere

by Vaclav Smil

Scientific American Library/W.H. Freeman: 1997. Pp. 221. \$32.95, £19.95

Mike Whitfield

If the whole of the Earth's history were compressed into a day, the residence times of most of the essential elements in the atmosphere, oceans and biosphere would be measured in fractions of a second. Only by parsimonious recycling and careful reutilization of elements could life persist and evolve in the face of such dynamic change. Smil's book addresses the biological cycling of just three elements — carbon, nitrogen and sulphur — which are essential to life and are effectively recycled in soluble waterborne compounds as well as in gases.

When dealing with terrestrial and atmospheric cycles, the book is authoritative and detailed, and is a pleasure to read. The way the cycles work, the evolution of our understanding of their intricacies and the interfer-

ence of civilization in their functioning are traced in clear language, and with beautifully presented diagrams and photographs. Nicely drawn dioramas or complex and intricate flowcharts illustrate the global cycles themselves. Frustratingly, fluxes and reservoir sizes are given only on subsets (Antarctic carbon cycle or Chinese agricultural nitrogen cycle) that are atypical of the global system.

A readily digestible analysis is provided of such complex issues as soil and water acidification and the fertilization of terrestrial plant growth by industrially released carbon dioxide and oxides of nitrogen. The folly of reaching for technofixes to alleviate such problems is emphasized most tellingly in an account of the intensification of the acid-rain problem in the United States that resulted from concerted efforts to clean up power-station discharges.

The author is less sure footed when dealing with the marine system. This is unfortunate because the oceans cover two-thirds of the Earth's surface, and for the first eighty per cent of the Earth's history provided the only environment of any significance. The role of the oceans in the carbon cycle is given fullest coverage, though even here the treatment is

oddly skewed at times. For example, the focus on the Antarctic carbon cycle leaves the false impression that, globally, birds and mammals form the main route for the transfer of carbon dioxide from the oceans to the atmosphere, and that only ten per cent of the carbon dioxide fixed by phytoplankton passes through the bacterial system.

The marine cycling of nitrogen gets short shrift. It is not included in a neatly presented history of our understanding of the nitrogen cycle, and the only mention within the chapter called "Reactive Nitrogen in the Biosphere" is an unexplained link between nitrate discharge and the crown-of-thorns starfish plague on the Great Barrier Reef. The production of dimethyl sulphide in the oceans is mentioned, but there is no overall budget for the global sulphur cycle to allow its impact to be assessed. Furthermore, its involvement in influencing cloud albedo over the open oceans, a matter of intense study and debate, is dismissed as unimportant, even in the preindustrial world, for the rather odd reason that burning vegetation also generates cloud-condensation nuclei. These criticisms are alleviated to some degree by an excellent treatment of the prob-

In retrospect chosen by Keith Stewart Thomson

Natural Theology; or Evidences of the Existence and Attributes of the Deity Collected from the Appearances of Nature

by the Reverend William Paley
(1802)

The alternative concepts to evolution are teleology and the argument from design, which is the proposition that the intricacies of nature and the supposed perfection of an organism's adaptation to the conditions of its existence (a darwinian term) are evidence of the operation of a First Cause, God the Creator. Paley's classic exposition of this subject has enjoyed almost as many editions and readers as Darwin's *On the Origin of Species*. Not that the argument from design was original to Paley; indeed a contemporary reviewer even asked why another work on this subject was necessary (*Edinburgh Review*, vol. 1, 1802–1803). As an undergraduate, Darwin himself was impressed by Paley's logic, although not unconditionally ("I did not at that time trouble myself about Paley's premises").

Natural theology is not so much a religious argument against the rising sciences of the early nineteenth century as a marshalling of scientific evidence against various forms of atheistic materialism. Paley's targets are David Hume, unitarians, deists and other free-thinkers including Buffon and, of course, Erasmus Darwin, and all notions that the natural world is the product of chance or the blind workings of laws and principles.

The trouble with Paley's argument is not its

form, and certainly not its brilliant opening syllogism: a watch is found on the ground, "we perceive... that its several parts are framed and put together for a purpose...(and) the inference, we think, is inevitable, that the watch must have had a maker." Therefore complex organisms must also have had a maker. (A similar argument was made by Cicero in *De Natura Deorum*.)

Paley must next imagine that a watch could be found that contained within it the makings of more watches ("... a complex adjustment of lathes, files, and other tools evidently and separately calculated for this purpose"), and now the downhill slide begins. Just as a hypothesis that explains everything explains nothing, so a Creator who has created everything has created too much. The First Cause cannot have unintended consequences, so Paley is left wriggling on his self-imposed hook: why did God create pain and suffering, war, pestilence and famine? (For Darwin, the heartbreaking death of his daughter Annie turned him against religion as much as any science.)

Paley, writing as a scientist rather than a theologian, takes some of his answers from Thomas Malthus (*An Essay on the Principle of Population*, 1798). For example, the malthusian inevitability of poverty "necessarily imposes labour, servitude, restraint". This natural cause of class structure forms a convenient argument against all kinds of disturbing movements of self-improvement. Where Malthus argued that utopian communities were impossible, Paley has to argue that we (blessed are the poor, indeed) live in one.

Malthus explains the harsh exigencies of the natural world, and Paley's works (which Darwin studied intently and partially memorized) thus gave Darwin an early (and unacknowledged) exposure to Malthus on population and key arguments such as 'superfecundity' in nature and the struggle for existence ("that system of natural hostilities").

Paley's complaints about Erasmus Darwin's transformism are familiar: "... a total lack of evidence. No changes, like those which the theory requires, have ever been observed." However, in his incisive articulations of the arguments for the Creator, Paley (like Lyell later) lays out a template which, if inverted, becomes a programme for the investigation of transmutation.

These remarkable passages were intended as irony: "... A countless variety of animals might have existed, which do not exist... unicorns and mermaids, sylphs and centaurs... nations of human beings, with more or fewer fingers and toes than ten... we may modify any one species many different ways, all consistent with life, and with actions necessary to preservation, although affording different degrees of convenience and enjoyment to the animal..." Imagine this being read by Robert Chambers, whose *Vestiges of the Natural History of Creation* (1844) in many ways broke ground for Darwin, and who was hexadactyl!

Keith Stewart Thomson is University Distinguished Scientist in Residence at the New School for Social Research, 65 West 11th Street, New York, New York 10011, USA.

lems involved in quantifying the impact of man on the global cycles.

The terrestrial bias may be excused in part because the immediate impact of man is most apparent on land. However, by inadvertently relegating the oceans to a secondary role, many of the feedbacks that are crucial to the stability of the biospheric cycles over the medium term are not considered. Indeed, the book deals unconvincingly with the overall potential for feedback in the natural cycles despite Smil's recognition of the "endless opportunities for interaction within and among individual cycles". The insights and provocative analyses of Lovelock in relation to feedback and regulation in the biosphere receive no mention, even though the early contributions of Vernadsky are recognized.

The book closes with an enigmatic quotation from Vernadsky in which he suggests that we are "in a transition to the noosphere", a concept related by Smil, in another quotation from Vernadsky, to "the reconstruction of the biosphere (my italics) in the interests of freely thinking humanity as a single totality". After reading this book I recoil from this vision! We should rather understand our dependence on these grand cycles, and learn to tread softly in their presence. □

Mike Whiffeld is Director of the Marine Biological Association of the United Kingdom, The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK.

Microbes writ large

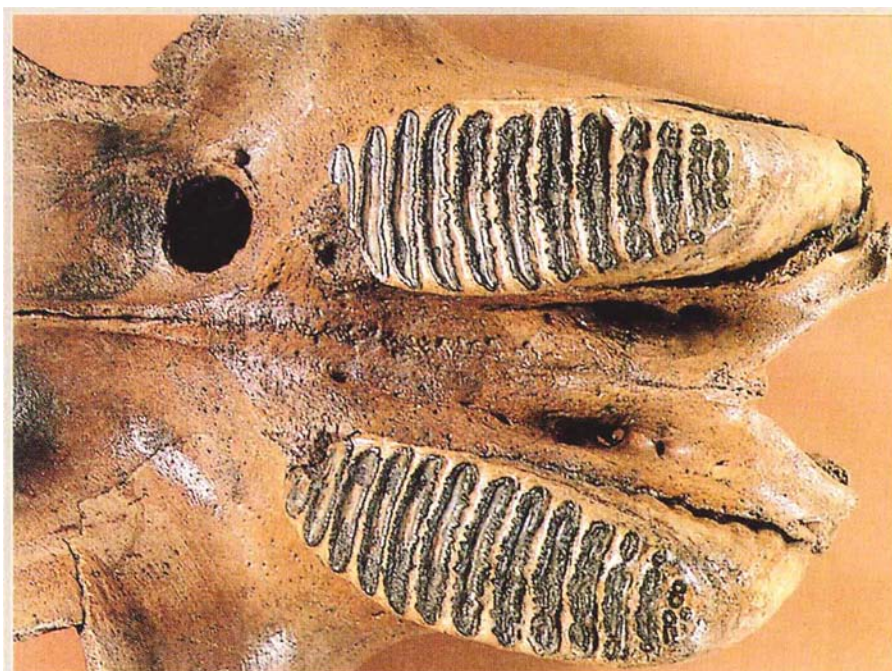
Life at Small Scale: The Behavior of Microbes

by David B. Dusenbery
Scientific American Library/W. H. Freeman:
1996. Pp. 214. \$32.95, £19.95

Roger Whittenbury

"A word means what I want it to mean" would (with apologies to Lewis Carroll) be an apt subtitle for this book. Imprecise as the term 'microbe' is, most microbiologists — indeed biologists — would agree that at the very least a microscope rather than a magnifying glass is required to see such a creature. The author does not subscribe to this view. Idiosyncratically, he opts for a functional definition: that microbes are organisms small enough to prosper without a circulatory system. This novel definition allows him to include in his 'microbial net' beasts such as nematodes and flatworms, and is an early indication of his unconventional approach. It is, in fact, hard to say exactly what message he is seeking to convey. The book comes across largely as a compilation of the author's interests, with readers left to get from it whatever they can.

The main themes explored are how microbes move, navigate and communicate, and there are diversions into survival and



Even elephants get toothache

This fossilized jaw of *Phalaeoloxodon*, an elephant from the Pleistocene epoch, has a circular hole by the left molar caused by

tooth decay. Helmut Mayr's *A Guide to Fossils* covers a wide range of fossils, and is out now in paperback (Princeton University Press, \$18.95).

feeding strategies. The result is a curious but interesting mix of detailed analyses, often with biophysical overtones (reflecting the author's background), basic commentary, and speculation. All this is accompanied by superb colour photographs and clear diagrams.

Ideally, readers should have at their side a standard text on microbiology, because the author makes generalizations that sometimes are not as all-embracing or accurate as they might have been. For instance, not all photosynthetic organisms are aerobic photoautotrophs that synthesize macromolecules from inorganic nutrients, carbon dioxide and water, releasing oxygen as a by-product. Many can photosynthesize only in the absence of oxygen, using a range of inorganic and organic electron donors other than water.

Occasional contradictory statements and apparent ambivalence on the author's part may be off-putting for the nonspecialist reader. The author says that *Synechococcus*, a photosynthetic cyanobacterium, produces oxygen in the dark and fixes nitrogen in the light, but on the next page we are told (correctly) the opposite. He also states that bacteria are too small to detect gradients of nutrients, suggesting that they have no choice but to settle for the niche in which they find themselves. Yet later on there is an extensive and excellent essay on chemotaxis, the process whereby motile bacteria can move towards or away from substances — in effect, up and down gradients. Although sensing a gradient along its surface might not be a property of a particular bacterium, the

author's treatment will undoubtedly puzzle some readers.

The ability of photosynthetic organisms to coexist despite their different physiological requirements is a fascinating story of evolutionary adaptation, tailor-made for a text of this type. Sadly, by presenting the facts too briefly and too simply, the author misses this opportunity. The description of chemotrophs provides another example of the dangers of oversimplification. The author gives the impression that all such organisms are strict anaerobes. In fact, most are strict aerobes that play an essential part in global nutrient recycling, using oxygen as an electron acceptor in the oxidation of a wide range of inorganic compounds such as ammonia, nitrite and ferrous sulphide.

The chapter on "Locomotion without legs" is a *tour de force*, however, heavily underpinned by explanations of the way in which physical principles dictate movement in water and how small organisms have adapted to viscosity and other physical constraints. Other chapters are more variable in content and detail, and less focused.

Overall, this is an interesting but not unflawed work, filling no particular niche. The intelligent layperson will find many parts difficult to comprehend but will be charmed by others. The author clearly is not troubled by such issues. He dedicates the book "To all who share in the wonder of life on Earth"! □

Roger Whittenbury is in the Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK.

The physics of microwave background anisotropies

Wayne Hu, Naoshi Sugiyama & Joseph Silk

The residual radiation from the Big Bang is observable as the cosmic microwave background—a near-uniform sea of photons with an average temperature of 2.7 K. Small variations in the temperature of the microwave background reflect the conditions that prevailed in the early Universe. Analysis of the physical mechanisms by which these anisotropies formed shows how they may be used to constrain fundamental cosmological parameters and provide insight into the origin of large-scale structure in the Universe.

Observations of the cosmic microwave background (CMB), a thermal sea of 2.7 K photons, form the cornerstone of modern cosmology. The very presence of the CMB provides strong support for the expanding hot Big Bang paradigm of cosmology¹. Small variations in its temperature across the sky (anisotropies) at the level of 10^{-5} encode a wealth of cosmological information that can be extracted from a detailed understanding of their formation.

Anisotropies in the CMB are so useful in cosmology because they originated during a much simpler epoch in the evolution of the Universe. CMB photons were released from ordinary matter a few hundred thousand years after the Big Bang as the Universe cooled to the point at which atomic hydrogen could exist. The physical processes involved in anisotropy formation are well understood, predictable and employ relatively few ingredients: gravity, thermodynamics and fluid dynamics. Their small amplitude implies that CMB anisotropies result from simple linear processing of the primordial fluctuations through gravitational and particle interactions. These same fluctuations grow by gravitational attraction into the large-scale structures observable in the Universe today.

The primordial fluctuations themselves may have their origin in a period of 'inflation' or rapid expansion which carried quantum-mechanical fluctuations to cosmological scales. Alternatively, a phase transition in the early Universe may have left topological defects analogous to domain walls in a ferromagnet. From observations of the anisotropy, one can thus extract information not only on the state of the Universe at an earlier time, including the parameters that describe the cosmological fluids and the dynamics of the expansion, but also on the origin of perturbations in the early Universe. Such inferences have significant implications for theoretical particle physics.

Large-angle anisotropies have been definitely measured near and above the 10° scale by the COBE Differential Microwave Radiometer (DMR) experiment². However, it is at the degree and subdegree scales that most of the cosmological information lies. As light can only travel a finite distance from the time at which the fluctuations formed, scales separated by more than this so-called 'horizon' could not have established contact in the interim. The horizon during anisotropy formation subtends no more than a degree on the sky today under the usual cosmological assumptions. Under this scale, the motion of matter generates sound waves³⁻⁵ with a harmonic structure that encodes a wealth of cosmological information, including the curvature, ordinary (baryonic) matter content, dark matter content, cosmological constant and the origins of perturbations in the Universe.

Although over the past four years great progress has been made in detecting degree- and subdegree-scale fluctuations⁶, systematic errors, atmospheric noise, extragalactic sources and Galactic emissions have prevented definitive measurements at these scales. With their large sky and frequency coverage coupled with recent advances in detector technology, experiments at present in the

planning stage (most notably two new satellite missions: see last section) are expected to overcome such difficulties within the next few years.

Cosmological framework

The cooling of CMB photons due to the cosmological expansion implies that in the early Universe when the CMB temperature is $T \geq 3,000$ K, the photons are hot enough to ionize hydrogen. During this epoch, the electrons 'glue' the baryons to the photons by Compton scattering and electromagnetic interactions. The dynamics that result involve a single photon-baryon fluid. Primordial perturbations gravitationally attract this fluid and try to compress it in the potential wells of high-density regions. Photon pressure resists this compression and sets up sound waves or acoustic oscillations in the fluid.

As the temperature drops below 3,000 K, the free electrons which trap the photons vanish as atomic hydrogen forms. Regions of compression and rarefaction at this epoch, called last scattering,

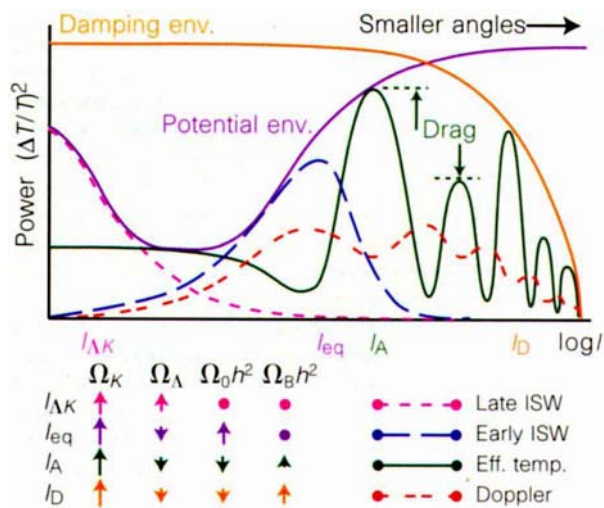


Figure 1 Schematic decomposition of the anisotropy spectrum and its dependence on cosmological parameters, in an adiabatic model. Four fundamental angular scales characterized by the angular wavenumber $l \propto \theta^{-1}$ enter the spectrum: l_K and l_Λ which enclose the Sachs-Wolfe plateau in the potential envelope, l_0 the acoustic spacing, and l_B the diffusion damping scale. The inset table shows the dependence of these angular scales on four fundamental cosmological parameters: $\Omega_K (= 1 - \Omega_\Lambda - \Omega_0)$, Ω_Λ , Ω_0/h^2 and Ω_B/h^2 (see Box 1 for definitions). Baryon drag enhances all compressional (here, odd) maxima of the acoustic oscillation, and can probe the spectrum of fluctuations at last scattering and/or Ω_B/h^2 . Projection effects smooth Doppler more than effective-temperature features.

Box 1 Mathematical description

Cosmological parameters. Cosmological densities today, for example ρ_X for the “X” constituent, are generally described in units of the critical density $\Omega_X = \rho_X/\rho_c$ where $\rho_c = 3H_0^2/8\pi G$ with the Hubble constant, which controls the present expansion rate, written as $H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1}$. Physical densities are thus proportional to $\Omega_X h^2$ and such quantities often appear in the text. Because the CMB fixes the radiation density under the standard thermal history, the matter-radiation ρ_m/ρ_r and baryon-photon ρ_B/ρ_γ , density ratios that control the dynamics are proportional to $\Omega_X h^2$ and $\Omega_B h^2$, respectively. The same notation applies to the vacuum density contribution ρ_Λ or cosmological constant $\Omega_\Lambda = \rho_\Lambda/\rho_c$. Furthermore, it is often convenient to express the curvature in units of the critical density $\Omega_K = -(c/H_0)^2 K$ because of the constraint $\Omega_K + \Omega_0 + \Omega_\Lambda = 1$.

Normal modes. A general spatial fluctuation (in, for example, the temperature or gravitational potential) may be conveniently represented as a superposition of normal modes that each evolve independently. In flat space, this corresponds to a Fourier decomposition of the fluctuation into plane waves of co-moving wavenumber k . We often consider the evolution of fluctuations in a single mode with the understanding that the observed anisotropy results by superposition of such modes with weights depending on the spectrum of primordial perturbations.

Oscillator dynamics. An isotropic temperature perturbation $\Theta = \Delta T/T$ in mode k evolves almost as a simple harmonic oscillator before last scattering¹⁴, $m_{\text{eff}}\ddot{\Theta} + k^2 c^2 \Theta/3 \approx m_{\text{eff}} g$. The overdots represent derivatives with respect to conformal time τ , which scales out the expansion, and the effective dimensionless mass of the oscillator is $m_{\text{eff}} = 1 + R$. Here $R = 3\rho_B/4\rho_\gamma$ is the baryon-to-photon momentum density ratio in the fluid. The oscillation frequency obeys the dispersion relation $\omega = kc/\sqrt{3m_{\text{eff}}} = kc_s$, where c_s is the sound speed. Gravity provides an effective acceleration of $g = -k^2 c^2 \Psi/3 - \dot{\Phi}$, where Ψ is the newtonian gravitational potential and $\Phi \approx -\Psi$ is the curvature perturbation on spatial hypersurfaces. These are related to the density fluctuation via the Poisson equation and give rise to gravitational infall and redshift from ‘metric stretching’ (see section in text on the integrated Sachs–Wolfe effect). The phase of the oscillation is given by $\phi = \int \omega d\tau = ks$ where the sound horizon $s = \int c_s d\tau$ is the distance sound can travel in time τ . This is the fundamental scale of acoustic oscillations.

If the potentials and baryon contribution R were constant, the fluid oscillates as $\Theta + \Psi = \frac{1}{3}\Psi(1+3R)\cos(ks) - R\Psi$, with compressional peaks a factor of $(1+6R)$ larger than the $\frac{1}{3}\Psi$ Sachs–Wolfe plateau and a difference in peak amplitude of $2R\Psi$ between even and odd peaks. The Doppler effect due to the line-of-sight velocity causes an r.m.s. temperature shift from $V_\parallel = (1/\sqrt{3})v_\parallel/c$, where $v_\parallel = -3\dot{\Theta}/k$ is the fluid velocity. Doppler oscillations $V_\parallel = \frac{1}{3}\Psi(1+3R)(m_{\text{eff}})^{-1/2}\sin(ks)$ are suppressed compared to temperature oscillations due to conservation of energy $\frac{1}{2}m_{\text{eff}}V_\parallel^2$. As the effective mass increases the velocity of the fluid decreases so that its contribution is suppressed by a factor of $m_{\text{eff}}^{-1/2}$.

In reality, these effects are modified due to the time evolution (see section in text on driving effects). One such complication is that as the effective mass evolves so does the amplitude. In classical mechanics, the ratio of energy $\frac{1}{2}m_{\text{eff}}\omega^2 A^2$ to frequency ω of an oscillator is an adiabatic invariant. Thus for the slow changes in $m_{\text{eff}} \propto \omega^{-2}$, the amplitude of the oscillation varies as $A \propto m_{\text{eff}}^{-1/4} \propto (1+R)^{-1/4}$ representing a small decay with time.

Projection geometry. The angular-size d relates the physical size of features λ_{feature} on a distant surface, say the last scattering surface, to their angular extent on the sky. It preserves the euclidean relation $\theta_{\text{feature}} \approx \lambda_{\text{feature}}/d$ in radians. Translated into wavenumbers, the relation is $l_{\text{feature}} \approx k_{\text{feature}}d$. In a curved universe, the angular-size distance from the present time t_0 to the surface at τ_* is not simply the ordinary distance $c(\tau_0 - \tau_*)$ but rather $d = |K|^{-1/2} \sinh[|K|^{1/2} c(\tau_0 - \tau_*)]$, where $K < 0$ is the curvature (for $K > 0$, $\sinh \rightarrow \sin$). The curvature enters (as described in the text) because gravitational lensing by the background magnifies or demagnifies the feature.

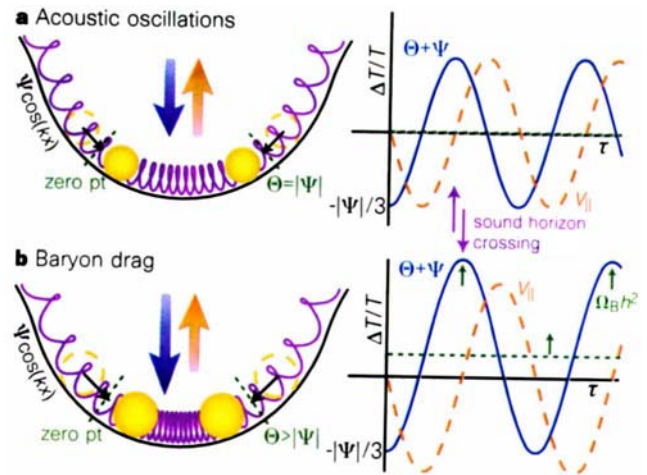


Figure 2 a, Acoustic oscillations. Photon pressure resists gravitational compression of the fluid setting up acoustic oscillations (left panel, real space). Spring and balls schematically represent fluid pressure and effective mass, respectively. Gravity displaces the zero point to $|\Psi|$ (blue arrow) with $\Psi/3$ oscillations (right panel, time). The displacement is cancelled by the redshift (red arrow) a photon experiences climbing out of the well. The Doppler effect ($V_\parallel$, red dashed lines) is shifted by $\pi/2$ in phase from the effective temperature ($\theta + \Psi$, blue solid lines). **b**, Baryon drag decreases the sound horizon and increases the gravitating mass, causing more infall and a net zero-point displacement, even after redshift (unequal red and blue arrows). Temperature crests (compression) are enhanced over troughs (rarefaction) and Doppler contributions.

represent hot and cold spots respectively in the CMB. The photons also suffer gravitational redshifts from climbing out of the potentials on the last scattering surface and from any time variation in the potentials along their trajectories. The combination of the two effects results in temperature differences on the last scattering surface which appear to the observer today as temperature anisotropies on the sky. We will now examine in greater detail the physical mechanisms that are responsible for the observable anisotropies schematically displayed in Fig. 1 (see refs 7–12 for more quantitative treatments).

Seeing sound

The gravitational infall of the fluid into potential wells generates compressional or acoustic waves. Because compression raises the temperature, this results in temperature oscillations that are visible on the sky today. We begin here with the basic picture of their formation and add complications that embed useful cosmological information in the following sections.

Let us first consider the effect of a static potential Ψ on the temperature perturbation $\Theta = \Delta T/T$ in a normal mode (see Box 1 and refs 13 and 14 for details). Though instructive, this approximation must be supplemented by evolutionary effects in any realistic model (see driving effects). For now, let us also ignore the dynamical effects of the baryons in the fluid.

Gravitational infall compresses the fluid until resistance from photon pressure reverses the motion. As a constant gravitational force merely shifts the zero point of the oscillation to $\Theta = -\Psi$, the initial displacement $\frac{1}{3}\Psi$ (Box 2) from the zero point oscillates as $\Theta + \Psi = \frac{1}{3}\Psi \cos(ks)$. Here k is the wavenumber and characterizes the length scale of the potential wells $\lambda = 2\pi/k$; s is the distance sound can travel by time τ , known as the sound horizon (Box 1 and Fig. 2a). Larger scales (lower k) oscillate more slowly owing to the finite time sound takes to cross the potential fluctuation. At a given time, measured by the size of the sound horizon s , different scales will be caught at different phases ϕ of their oscillation ($\phi = ks$).

After last scattering, the photons suffer a gravitational redshift Ψ from climbing out of the potential. We thus call $\Theta + \Psi$ the effective temperature fluctuation. Notice that the effective temperature oscillates around zero because gravitational infall and redshift have the same physical origin if the baryons are dynamically insignificant. We shall see that its evolution determines the structure of the small-angle spectrum (Fig. 1).

Acoustic peaks

At last scattering, the acoustic oscillations are frozen, leaving peaks where they are caught at extrema of their oscillation. These peaks form a harmonic series in wavenumber based on the sound horizon at last scattering and provide a scale through which quantitative measurements can be made.

The wavenumber that reaches its first compression by last scattering is $k_A = \pi/s_*$, and corresponds to the sound horizon at that time, which we denote here and below with a subscripted asterisk. Longer wavelengths will not have evolved significantly, yielding an effective temperature fluctuation of $\frac{1}{3}\Psi$ which directly reflects the primordial potential perturbations. This combination of the intrinsic temperature fluctuation and the gravitational redshift is the well known Sachs–Wolfe effect¹⁵ observed by the COBE DMR instrument. The phase evolves for shorter wavelength fluctuations. As a function of k , there will be a harmonic series of temperature fluctuation peaks with $k_m = mk_A = m\pi/s_*$ for the m th peak.

The series is based on the sound horizon s_* . Like the horizon, it is most sensitive to the expansion rate of the Universe before last scattering which in turn depends primarily on the ratio of matter to radiation in the Universe. Because the latter is fixed by the CMB itself, the sound horizon is determined mainly by the matter density today $\Omega_0 h^2$ (see Box 1 for notation).

These effective temperature peaks dominate the small-angle spectrum (Fig. 1). Odd peaks here represent the compression phase (temperature crests), whereas even peaks represent the rarefaction phase (temperature troughs), inside the potential wells.

Projection effects

The angular scale on which anisotropies appear on the sky depends on the geometry of the Universe, and the distance to the surface of last scattering of the microwave background photons. Physical features on that surface, such as those produced by the acoustic peaks, appear as angular features on the sky. By combining measurements of the physical and angular sizes of features, the curvature of the Universe can be determined from the CMB.

Consider first the case of positive curvature. Photons stream to the observer on geodesics, analogous to lines of longitude to the pole (Fig. 3). Positive curvature (closed universe) makes the same physical scale at fixed latitude (distance) subtend a larger angle than in the euclidean case. The opposite effect occurs for negative curvature (open universe)^{13,16}. This can also be thought of as the angular magnification by gravitational lensing from the background density.

Increasing the distance to the last scattering surface also decreases the angular extent of acoustic features. This distance depends mainly on the expansion rate and hence the matter content (Box 1). Because the same quantity enters length scales such as the sound horizon s_* (see previous section), this dependence nearly cancels leaving the much stronger dependence on the curvature. Putting these quantities together, we can construct an angular-size distance d which depends on the curvature and is defined to follow the euclidean expectation that $l_{\text{feature}} = k_{\text{feature}} d$. Here the angular wavenumber l describes the angular extent of the feature, $\theta \approx 100/l$ degrees. Features, especially the peak spacing l_A based on k_A , provide angular-size distance measures of the curvature^{17,18}.

One complication arises. The way a feature appears depends on its orientation with respect to the last scattering surface, hitherto assumed face-on. Sharp features in k -space, such as zeros in the

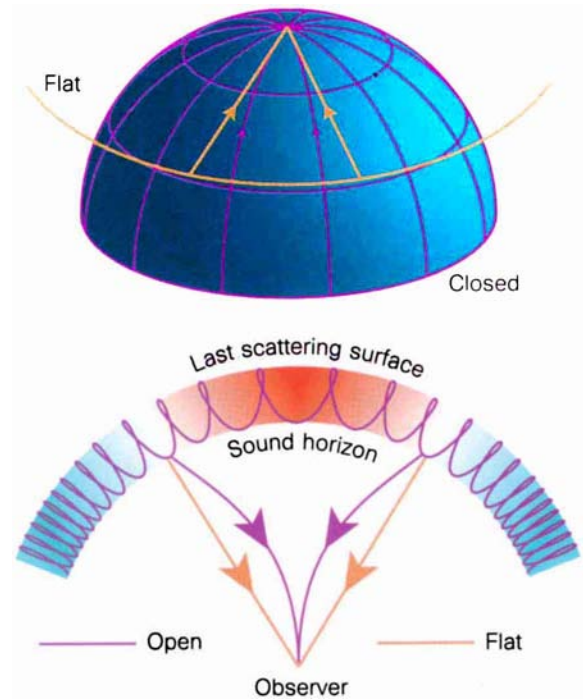


Figure 3 Projection effects. Temperature fluctuations on the last scattering surface appear as anisotropies on the sky. The angular size depends on the geometry of the Universe and the distance to the surface. At a fixed distance, a smaller physical scale is required to subtend the same angle in a closed universe (upper panel) and larger in an open universe (lower panel, schematically flattened for clarity). Physical scales such as the sound horizon at last scattering can be used to measure the curvature.

acoustic oscillation, are somewhat ‘washed-out’ in l -space as all orientations are superimposed (Fig. 1). Similarly, gravitational lensing by large-scale structure causes perturbations in the projection that can further wash out features in the anisotropy spectrum^{19,20}. Neither effect is sufficient to destroy the observability of the peaks entirely.

In summary, projection effects scale the features multiplicatively in angle with curvature and smooth them out somewhat. The combination of acoustic oscillations and projection effects provides the basic structure of CMB anisotropies.

Baryon drag

Baryons alter the dynamics of the oscillations because they have mass and are attracted into potential wells. The baryon content of the early Universe, as well as the amplitude of the potential at last scattering, can thus be measured through the heights of the peaks in the acoustic oscillations.

The baryon contribution to the effective mass of the fluid lowers the frequency of the oscillation, but more importantly, also causes greater gravitational compression of the fluid in a potential well, that is, a further displacement of the oscillation zero point¹⁴ (see Fig. 2b and Box 1 for further details). The same initial conditions represent larger-amplitude oscillations around the zero point. Moreover, we see the shift directly in the power spectrum (squared temperature fluctuation) as alternating peak heights. Peaks from compression (here odd) are enhanced over those from rarefaction with increasing baryon content and depth of potential wells (Box 1 and Fig. 1).

Doppler shifts

The motion of the fluid Doppler-shifts the frequency of the photons, imprinting oscillations which are less prominent than—

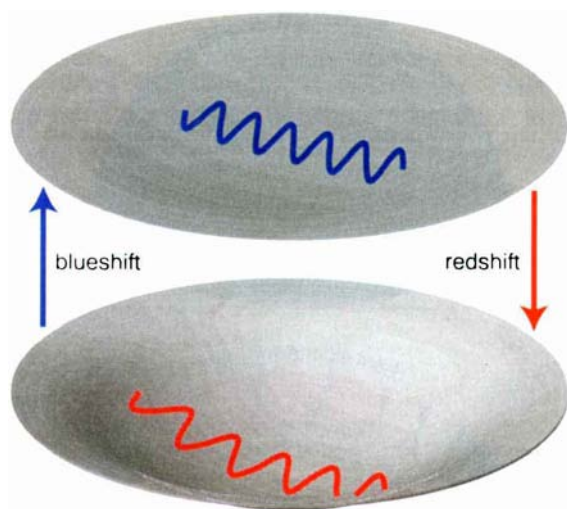


Figure 4 Metric stretching. Changes in the spatial metric distort the wavelength of passing photons. If the potential decays, the contracting curvature perturbation blueshifts the photons (bottom to top). If the potential grows, the corresponding 'stretching' of space redshifts the photons (top to bottom). This effect doubles the ordinary potential infall-redshift effect for both the driving force to the acoustic oscillator and the ISW effects.

and out of phase with—the intrinsic temperature oscillations. The Doppler effect adds a smoother component to the anisotropy that makes the acoustic peaks rather less prominent for low-baryon-content universes (see Fig. 1). The common use of the term 'Doppler peak' to refer to the acoustic peaks in the power spectrum is a misleading historical accident.

As the turning points of the oscillation are at the temperature extrema, the fluid velocity oscillates $\pi/2$ out of phase with the temperature (Fig. 2a). Line-of-sight motion relative to the observer causes a Doppler shift. Whereas the observer velocity creates a pure dipole anisotropy on the sky, the fluid velocity causes an r.m.s. spatial temperature variation on the last scattering surface from its line-of-sight component⁵. In the absence of baryons, the velocity contribution is equal in amplitude to the temperature effect. But because the line-of-sight restriction eliminates the face-on contributions depicted in Fig. 3, projection effects smooth Doppler features more strongly than temperature features¹⁴. This implies that the peak structure of the anisotropy is due mainly to the intrinsic temperature.

Baryons increase the prominence of the peaks by decreasing the relative contribution of the Doppler effect. As the baryons increase the effective mass, conservation of energy requires that the velocity decreases as the inverse square-root of the mass for the same initial temperature displacement (Box 1). We note that velocity oscillations are symmetric around zero making the baryon drag effect even more prominent (Fig. 2a).

Driving effects

All realistic models of the early Universe involve time-dependent potentials causing forced oscillations in the baryon–photon fluid (Fig. 4). The amplitude and phase of the oscillations records information regarding how the potential evolves in time.

By assuming that the potentials are constant, we have hitherto implicitly assumed that matter dominates the energy density. In reality, radiation dominates early on. We show in Box 2 how radiation fluctuations introduce a distinct class of isocurvature perturbations where pressure plays a dynamic role in potential evolution.

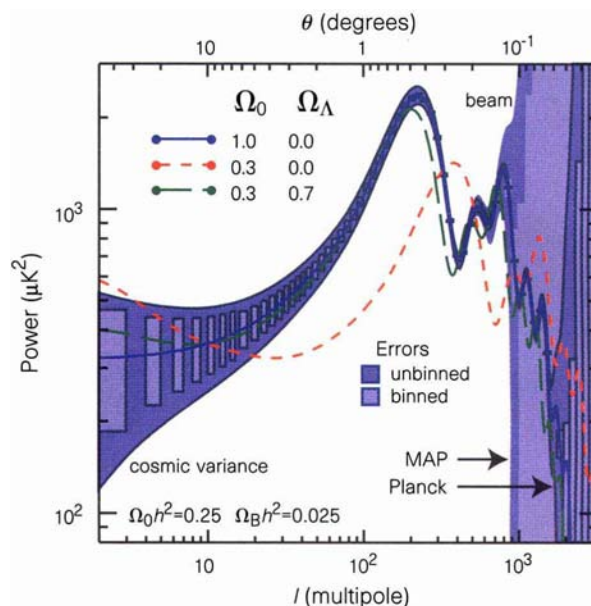


Figure 5 Idealized performances of two new satellites with full sky coverage (see text): MAP and Planck Surveyor. An underlying $\Omega_0 = 1$ inflationary spectrum (dark blue curve) can be recovered within the blue error contour (dark blue outline for C/S). At large angles ($\theta \approx 100^\circ$ degrees), deviations are due to 'cosmic variance', which is the same for both satellite experiments; at small angles, deviations are due to the different beam sizes and sensitivities of the two satellites. Binning reduces the recovery errors on the smooth underlying power spectrum (light blue boxes). The $\Omega_0 = 0.3$ open (red short-dashed lines) and even the Ω_Λ (green long-dashed lines) variant are easily distinguished under the scale-invariant inflationary paradigm. More robust parameter measurements require that several acoustic features be resolved by the Planck Surveyor. In practice, foreground contamination will reduce the amount of sky available, increasing the errors as $f_{\text{sky}}^{-1/2}$, foreground subtraction will increase the noise, and secondary anisotropies will distort the spectrum at high l .

Cosmological defect models are part of this class of isocurvature perturbations. Indeed, because inflation is the unique causal generator of adiabatic fluctuations, observational distinctions between adiabatic and isocurvature models are important for particle physics¹⁹. As discussed further in Box 2, the driving mechanism of isocurvature perturbations causes a phase shift in the acoustic oscillation $\cos(k_s + \phi)$ which is typically $\phi \approx \pi/2$. This phase shift can be measured by determining the locations of the peaks k_m . Thus by measuring at least two peaks, one can learn important properties of particle physics in the early Universe.

In both the adiabatic and isocurvature scenarios, driving effects, where the gravitational 'push' is timed to match the frequency of the oscillation, typically boost the amplitude of the oscillations and leave an imprint of the matter–radiation transition characterized by the potential envelope (see Fig. 1). The critical scale corresponds to the horizon at matter–radiation equality, $k_{\text{eq}} \propto \Omega_0 h^2$. For adiabatic fluctuations, acoustic oscillations smaller than this scale are boosted by a factor of ~ 5 above the Sachs–Wolfe plateau in temperature. The transition is traced by the potential envelope, and occurs at $l_{\text{eq}} = k_{\text{eq}} d$. The driving effect therefore introduces a second characteristic scale l_{eq} to the anisotropy spectrum through the potential envelope (Fig. 1) and sets the harmonic relation between the ratios of the acoustic peak locations.

Diffusion damping

Damping processes cut off the spectrum of acoustic oscillations at small angular scales, thereby imprinting a third physical scale—the

photon-diffusion scale at last scattering—into the CMB (see Fig. 1, where the power goes to zero at high l). Like the sound horizon, this scale depends only on the background cosmology, such as the thermal history and baryon content of the fluid. Measuring the diffusion damping scale l_D is therefore useful for extracting information about the thermal history and baryon content from the anisotropy power spectrum.

In reality, the photon–baryon fluid is imperfect because the photons random-walk through the baryons with a mean free path λ_C given by Compton scattering. As the photons diffuse, hot and cold regions mix, thereby cancelling out fluctuations. Fluctuations are damped nearly exponentially as the diffusion length $\lambda_D \approx \sqrt{N\lambda_C} = \sqrt{c\tau\lambda_C}$ overtakes the wavelength³. At last scattering, the ionization fraction x_e decreases due to recombination, thus increasing the mean free path of the photons $\lambda_C \propto x_e^{-1}$. The diffusion scale $k_D = \lambda_D/2\pi$, can therefore be used as a probe of the ionization history and the baryon content $\Omega_b h^2$.

Integrated Sachs–Wolfe effect

Photons experience gravitational blueshifts and redshifts when traversing a varying gravitational potential (or, more generally, perturbations in the geometry of space-time known as ‘metric’ fluctuations). This effect, as applied to the microwave background, is generally referred to as the integrated Sachs–Wolfe (ISW) effect¹⁵. As the evolution of structure changes the depth of potential wells, the redshift experienced by a photon as it leaves the well no longer exactly cancels the blueshift it encountered when entering, which leads to a temperature fluctuation. A second related effect doubles the result. As the Universe expands, the wavelength of a photon is redshifted. Similarly, the wavelength also stretches as the overdensities which create potential wells stretch the spatial fabric (Fig. 4). For example, as the potential decays, the contraction of the photon wavelength blueshifts the CMB to higher temperatures. This heuristic explanation also holds for more complicated metric fluctuations. This mechanism provides sensitivity to the equation of state, expansion dynamics and nonlinear clustering^{24–26}, through their effects on the evolution of structure, as well as to any external source of metric fluctuations such as gravitational waves^{21–23} and cosmological defects^{27–29}, for example, cosmic strings³⁰.

Early ISW effect. The early ISW effect arises if the Universe is not completely matter-dominated at last scattering¹⁴. For adiabatic models, the potential decays for modes that cross the sound horizon between last scattering and full matter-domination $k_{eq} \leq k \leq k_A$, yielding a net effect up to five times greater than the Sachs–Wolfe temperature plateau $k_{eq} \ll k_A$ (low $\Omega_0 h^2$). In isocurvature models, the potential growth outside the sound horizon can contribute more than the Sachs–Wolfe effect for such low $\Omega_0 h^2$.

Late ISW effect. In an open (K) or cosmological constant (Λ) model, the universe enters a rapid expansion phase once matter no longer dominates the expansion^{31,32}. The potential decays to zero within an expansion time, independent of the wavelength. For the adiabatic case, the maximum is again five times greater than the Sachs–Wolfe temperature plateau. But opposing effects from decaying overdensities and underdensities tend to cancel if the photon can travel across many wavelengths during the decay. Thus, this effect is suppressed below the horizon at the decay epoch τ_{KA} , $k \geq k_{KA} = 2\pi/\tau_{KA}c$ (ref. 33). As similar arguments hold for isocurvature models, a feature in the potential envelope is generally expected at $l_{KA} = k_{KA}d$ where d is now the angular-size distance to the epoch of decay.

These two manifestations of the ISW effect involve potential decay due to changes in the equation of state of the background from radiation- to matter-dominated, and from matter- to cosmological-constant or curvature-dominated. They imprint the scales l_{eq} and l_{KA} onto the anisotropy spectrum through the potential envelope (Fig. 1).

Secondary scattering

The first generation of stars and quasars can produce radiation to reionize hydrogen at late times. Subsequent rescattering both erases the primary anisotropies and generates new secondary ones. In models without high-amplitude small-scale fluctuations initially, such as the scale-invariant adiabatic model, reionization sufficiently early to reach high optical depth is unlikely. We therefore omit a detailed discussion of these effects³⁴.

Extracting information

The physical mechanisms described above encode information about cosmological parameters and the primordial origin of fluctuations into the CMB (Fig. 1). By understanding their effect on the power spectrum of anisotropies, we see how they may be determined from observations.

The four angular scales imprinted on the CMB, namely l_{KA} , l_{eq} , l_A and l_D , form the fundamental basis for the spectrum. We show a scale-invariant adiabatic model in Fig. 1. The first two scales appear because of the difference in potential evolution due to the equation of state of the Universe, that is, whether it is radiation-, matter-, curvature- or Λ -dominated. This defines a ‘potential envelope’ as shown in Fig. 1. The exact form of this envelope varies with the model but should generally bear the imprint of the two fundamental scales.

The adiabatic potential envelope separates cleanly into three basic pieces: the late ISW fall, the Sachs–Wolfe (SW) plateau, and the early-ISW/driven-oscillation rise. Acoustic effects also produce temperature as well as smaller, smoother Doppler oscillations regularly spaced by l_A . Photon diffusion provides a damping envelope with a characteristic scale l_D that is entirely independent of the initial conditions (Box 2).

For the standard thermal history and a universe composed of baryons, photons, cold dark matter, and massless neutrinos, these four scales can be combined to extract the four basic parameters^{33,35–37} Ω_K , Ω_Λ , $\Omega_0 h^2$ and $\Omega_b h^2$; these parameters are respectively the curvature, the cosmological constant, the matter content and the baryon content (see Box 1 for definitions and Fig. 1, inset table). Note that $\Omega_0 = 1 - \Omega_K - \Omega_\Lambda$ is not an independent parameter and hence the Hubble constant h may also be inferred from these measurements. We choose this representation as it is commonly supposed that either Ω_K or Ω_Λ vanishes and because the matter/radiation and baryon/photon density ratios control the physical processes in the early Universe.

Because l_A , the spacing between the acoustic peaks, is only weakly dependent on the other parameters and relies on the most distinctive features in the spectrum, it provides the cleanest measure of the curvature Ω_K (ref. 18). In a given model, the location of any peak is simply proportional to l_A and can alternately be used as the measure¹⁷. Even the degenerate but mild dependence on Ω_Λ can be removed by measurements of, or constraints on, l_{KA} . The dramatic change in the potential envelope at l_{eq} provides a sensitive probe of $\Omega_0 h^2$. In particular, l_{eq}/l_A is entirely independent of Ω_K and Ω_Λ and only weakly dependent on $\Omega_b h^2$ thus serving to isolate $\Omega_0 h^2$. One can also employ l_A/l_D , which is independent of Ω_K and Ω_Λ and weakly dependent on $\Omega_0 h^2$, to measure the baryon content.

Other properties of the spectrum can provide information on the primordial perturbations. The potential envelope measures its spectrum and possible tilts from scale invariance. The relative peak heights measure potentials at last scattering through the baryon drag mechanism. Alternatively, the effect can be used to measure $\Omega_b h^2$ in the context of a given model for the fluctuations. The relation between the m th peak location and the peak spacing l_m/l_A , or more generally the phase of the oscillation, reveals information about the driving mechanism and clearly separates inflationary models from most other causal models such as cosmological defects³⁸.

Box 2 Primordial fluctuations

The CMB also allows a relatively clean determination of the nature of the initial perturbations with perhaps far-reaching consequences for particle physics. The most fundamental distinction is the presence or absence of initial curvature perturbations. Causal curvature (adiabatic) perturbations require a period of superluminal expansion as in the inflationary model, leaving all other causal models in the isocurvature class. The acoustic peak locations and relative heights can distinguish between the two cases.

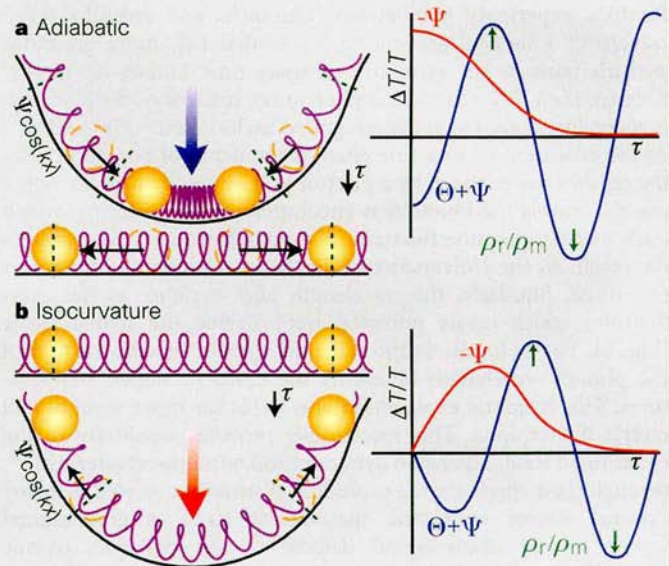
In the adiabatic case, number densities n of the different particle components initially fluctuate together. The resultant energy density perturbation δ generates a curvature and potential fluctuation $\Psi \approx -\Psi = \frac{1}{2}\delta$. If radiation dominates, $\delta = \delta_r = 4\Theta$ whereas if matter dominates $\delta = \delta_m = \delta n_r/n_r = 3\Theta$ where subscripts m and r refer to the matter and radiation, respectively. On the other hand, by balancing the initial fluctuations so that energy density perturbations cancel, isocurvature fluctuations may be established. Causality implies that the curvature perturbation remains constant until the matter can be redistributed. Thus in an isocurvature model, the gravitational perturbations only grow to be significant when the horizon grows larger than the wavelength, called 'horizon crossing'. Because they represent the force felt by the oscillator, adiabatic and isocurvature conditions yield distinctly different acoustic signatures. The details of how the isocurvature fluctuation is set up (for example, through balancing radiation fluctuations with baryons, cold dark matter, or defects such as textures⁴¹) is not as important as the fact that potential fluctuations vanish initially and then grow in anticorrelation with the photon fluctuations until horizon crossing¹⁸.

In both cases, the self-gravity associated with the acoustic perturbation can feed back into the potentials to drive the oscillator. For the adiabatic mode, the initial potential perturbation is formed in large part by the photon-baryon density fluctuation. At horizon crossing, the photon-baryon fluid begins to compress itself due to its self-gravity and the effective temperature reverses sign. As pressure tries to stop the compression, the potential decays. The fluid is left in a highly compressed state. Thus self-gravity acts as a driving term timed to enhance the first compression of a $\cos(k\tau)$ series (see box figure). This effect is doubled by dilation from the decaying curvature perturbation Φ (see Fig. 4). If self-gravity dominates, as in the case in which the universe was radiation-dominated at horizon crossing, the fluid is left in a cosine oscillation with $2\Psi - \frac{1}{3}\Psi = \frac{5}{3}\Psi$ or 5 times the amplitude of the Sachs-Wolfe tail. The m th peaks occurs at $k_m = mk_A$ and baryon drag enhances all odd peaks.

In the isocurvature case, photon density fluctuations are initially balanced by those of the other species to eliminate the curvature. Photon pressure intercedes near horizon crossing to break this balance, letting the potential fluctuation grow from zero. Still the photons attempt to compensate by becoming more and more underdense or rarefied inside potential wells $\Theta + \Psi \approx 2\Psi < 0$ (see also Fig. 4). This continues until sound horizon crossing where photon pressure successfully resists further rarefaction (see box figure). The fluid turns around and begins falling into the potential wells and furthermore enhances them by their self-gravity. As the photons resist further compression at the positive maximum, the self-gravity contribution to the potential fluctuation decays. This again leaves the photon-baryon fluid in a highly compressed state and increases the amplitude of the acoustic oscillation. Just as with the adiabatic case, the self-gravity of the photon-

baryon fluid essentially drives the oscillator. Unlike the adiabatic case, it drives the sine rather than the cosine oscillation. The m th peaks occur at $k_m = (m - 1/2)k_A$ and baryon drag enhances all even peaks⁴³. The harmonic series thus serves to distinguish the two cases through the peak location to separation $l_m/l_{\kappa} = k_m/k_A$.

Self-gravity is not necessarily the main forcing mechanism of isocurvature models. Strong features in the source potentials (for example, from nonlinear interactions in a cosmic string network) can change the phase shift and/or provide modulations of the peaks. Stochastic behaviour in such sources can also lead to randomness in the phase shift, which translates into a washing away of the peaks when k -modes are superimposed⁴². So under very general assumptions,²⁸ observation of the peak locations (which determine the phase shift and coherence) as well as the relative heights (which distinguishes between comparisons and rarefactions) allows a clean distinction between adiabatic and isocurvature models. This distinction provides potentially deep insight into the origins of the primordial perturbations.



Box figure Driven oscillations. **a**, Adiabatic case. As the fluid begins to compress (blue arrow), photon pressure resists the increase in the density perturbation thereby allowing the potential to decay. Left in a highly compressed state, the fluid then oscillates with enhanced amplitude. **b**, Isocurvature case. Radiation fluctuations are set up to eliminate the potential initially. Photon pressure resists the accompanying rarefaction (red arrow). The fluid then falls back into the well, lending its self-gravity to enhance the depth. At the compressional maxima, photon pressure again causes potential decay leaving the fluid in a highly compressed state. These feedback effects increase with the radiation-matter density ratio ρ_r/ρ_m driving a cosine oscillation in the adiabatic case and a sine oscillation in the isocurvature case.

Observational prospects

CMB anisotropy formation is a simple, well understood linear process. Its anisotropy spectrum thus provides some of the cleanest and most robust cosmological tests available. At present, extraction of cosmological parameters requires a specific model, but most of the underlying assumptions may be verified as the resolution reaches the several-arcminute range. Here the acoustic peaks reveal the contrasting signatures of, and provide consistency tests for, the various cosmological parameters and models.

To achieve these goals, coverage of a significant part of the sky is

also crucial. A large number of independent patches at any specified angular scale is needed to reduce the sampling variance. Minimal sampling variance is obtained with a full sky map, the so-called 'cosmic variance' limit (see Fig. 5 for an illustration using a calculated model spectrum). In practice, information from some fraction of the sky will have to be discarded owing to Galactic foregrounds emission. But because foregrounds such as Galactic synchrotron, bremsstrahlung and dust emission have different frequency and angular signatures, it should be possible to remove much of the contamination with multifrequency sky maps^{39,40}.

Another uncertainty lying ahead will be that of unresolved sources in the foreground. If at least half of the sky is mapped and relatively foreground free, it should be possible to measure most cosmological parameters at present being considered to a precision of better than a few per cent (ref. 37), as well as to determine the nature of the primordial fluctuations, with quality maps of 10–20 arcmin resolution.

These goals should be realized by two new satellites: MAP from NASA for launch in 2000 and Planck Surveyor from ESA with estimated launch in 2004. The raw specifications of the MAP mission include five frequency channels from 22 to 90 GHz with angular resolutions between 18 and 54 arcmin and a sensitivity of $35 \mu\text{K}$ per pixel per year (the pixel size is 18×18 arcmin). The Planck Surveyor mission should include nine frequency bands from 31.5 to 857 GHz and angular resolutions between 4.4 and 30 arcmin; the two central CMB channels (of pixel size 7×7 and 10×10 arcmin) have sensitivities of $3.3 \mu\text{K}$ and $5.5 \mu\text{K}$ per pixel per 14 months, respectively. How these specifications translate into properties of a CMB map will depend on systematic effects, foregrounds and so on. We provide a rough indication in Fig. 5 of the combined sensitivity of the MAP channels and a single channel of the Planck Surveyor. Though suggestive, it should not be taken literally as an assessment of their capabilities.

Within a decade from now, in the absence of unforeseen systematic effects and foregrounds, we will possess definitive maps of subdegree-scale anisotropies in the CMB. These measurements will undoubtedly match, and in many cases exceed, the precision of experimental particle physics measurements and correspondingly establish the cosmological model as securely as the Standard Model of elementary particles. We will then know as much, or even more, about the early Universe and its contents as we do about the fundamental constituents of matter. $\square$

W.H. is at the Institute for Advanced Study, School of Natural Sciences, Princeton, New Jersey 08540, USA. N.S. is at the Department of Physics, Kyoto University, Kyoto 606-01, Japan. J.S. is at the Department of Astronomy and Physics, and

Center for Particle Astrophysics, University of California, Berkeley, California 94720, USA.

1. Peebles, P. J. E., Schramm, D. N., Kron, R. G. & Turner, E. L. *Nature* **352**, 769–776 (1991).
2. Smoot, G. et al. *Astrophys. J.* **396**, L1–L4 (1992).
3. Silk, J. *Astrophys. J.* **151**, 459–471 (1968).
4. Peebles, P. J. E. & Yu, J. T. *Astrophys. J.* **162**, 815–836 (1970).
5. Sunyaev, R. A. & Zel'dovich, Ya. B. *Astrophys. Space Sci.* **7**, 3–19 (1970).
6. Scott, D., Silk, J. & White, M. *Science* **268**, 829–835 (1995).
7. Wilson, M. & Silk, J. *Astrophys. J.* **243**, 14–25 (1981).
8. Bond, J. R. & Efstathiou, G. *Astrophys. J.* **285**, L45–L48 (1984).
9. Vittorio, N. & Silk, J. *Astrophys. J.* **285**, L39–L43 (1984).
10. Hu, W., Scott, D., Sugiyama, N. & White, M. *Phys. Rev. D* **52**, 5498–5515 (1995).
11. Ma, C.-P. & Bertschinger, E. *Astrophys. J.* **455**, 7–25 (1995).
12. Seljak, U. & Zaldarriaga, M. *Astrophys. J.* **469**, 437–444 (1966).
13. Doroshkevich, A. G., Zel'dovich, Ya. B. & Sunyaev, R. A. *Sov. Astron.* **22**, 523–528 (1978).
14. Hu, W. & Sugiyama, N. *Astrophys. J.* **444**, 489–506 (1995).
15. Sachs, R. K. & Wolfe, A. M. *Astrophys. J.* **147**, 73–90 (1967).
16. Sugiyama, N. & Gouda, N. *Prog. Theor. Phys.* **88**, 803–844 (1992).
17. Kamionkowski, M., Spergel, D. N. & Sugiyama, N. *Astrophys. J.* **434**, L1–L4 (1994).
18. Hu, W. & White, M. *Astrophys. J.* **471**, 30–51 (1996).
19. Cole, S. & Efstathiou, G. *Mon. Not. R. Astron. Soc.* **239**, 195–200 (1989).
20. Seljak, U. *Astrophys. J.* **463**, 1–7 (1996).
21. Abbott, L. F. & Wise, M. B. *Nucl. Phys. B* **244**, 541–548 (1984).
22. Starobinski, A. A. *Sov. Astron. Lett.* **11**, 113 (1985).
23. Davis, R. et al. *Phys. Rev. Lett.* **69**, 1856–1859 (1992); **70**, 1733 (1993).
24. Rees, M. I. & Sciama, D. N. *Nature* **217**, 511–516 (1968).
25. Seljak, U. *Astrophys. J.* **460**, 549–555 (1996).
26. Tuluie, R. & Laguna, P. *Astrophys. J.* **445**, L73–L76 (1995).
27. Coulson, D., Ferreira, P., Graham, P. & Turok, N. *Nature* **368**, 27–31 (1994).
28. Pen, U. L., Spergel, D. N. & Turok, N. *Phys. Rev. D* **49**, 692–729 (1994).
29. Blanchard, A. & Schneider, J. *Astron. Astrophys.* **184**, 1–6 (1987).
30. Kaiser, N. & Stebbins, A. *Nature* **310**, 391–393 (1984).
31. Kofman, L. A. & Starobinski, A. A. *Sov. Astron. Lett.* **9**, 643–651 (1985).
32. Sugiyama, N. & Silk, J. *Phys. Rev. Lett.* **73**, 509–513 (1994).
33. Hu, W. & Sugiyama, N. *Phys. Rev. D* **51**, 2599–2630 (1995).
34. Bond, J. R. in *Cosmology and Large Scale Structure* (eds Schaffer, R. et al.) (Elsevier, in the press).
35. Bond, J. R. et al. *Phys. Rev. Lett.* **72**, 13–16 (1994).
36. Seljak, U. *Astrophys. J.* **435**, L87–L90 (1994).
37. Jungman, G., Kamionkowski, M., Kosowski, A. & Spergel, D. N. *Phys. Rev. D* **54**, 1332–1344 (1995).
38. Hu, W., Spergel, D. N. & White, M. astro-ph/9605193; *Phys. Rev. D* (in the press).
39. Brandt, W. N. et al. *Astrophys. J.* **424**, 1–21 (1994).
40. Tegmark, M. & Efstathiou, G. *Mon. Not. R. Astron. Soc.* **281**, 1297–1314 (1996).
41. Crittenden, R. & Turok, N. *Phys. Rev. Lett.* **75**, 2642–2645 (1995).
42. Magueijo, J., Albrecht, A., Coulson, D. & Ferreira, P. *Phys. Rev. Lett.* **76**, 2617–2620 (1996).

Acknowledgements. We thank E. Bunn, D. Scott and M. White for discussions. W.H. was supported by the US NSF and the W.M. Keck Foundation; J.S. was supported in part by NASA.

Correspondence should be addressed to W.H. (e-mail whu@ias.edu).

A Polycomb-group gene regulates homeotic gene expression in *Arabidopsis*

Justin Goodrich*†‡\$, Preeya Puangsomlee*†‡, Marta Martin†, Deborah Long†, Elliot M. Meyerowitz‡ & George Coupland†

† John Innes Centre for Plant Science Research, Colney Lane, Norwich NR4 7UH, UK

‡ Division of Biology 156-29, California Institute of Technology, Pasadena, California 91125, USA

\$ Present address: Institute of Cell and Molecular Biology, Rutherford building, The King's Buildings, Mayfield Road, Edinburgh EH11 1DN, UK

* These authors contributed equally to this work

Cell fate is determined when the commitment of cells to a particular fate is autonomously maintained, irrespective of their environment. In *Drosophila*, fate determination is maintained through the action of the Polycomb-group and trithorax-group genes, which are required so that states of homeotic gene activity are inherited through cell division. It is shown here that the *CURLY LEAF* gene of *Arabidopsis* is necessary for stable repression of a floral homeotic gene and encodes a protein with homology to the product of the Polycomb-group gene *Enhancer of zeste*. We suggest that Polycomb-group genes have a similar role in fate determination in plants and animals.

Homeotic genes specify the identity of serially repeated units such as the whorls of flowers or the segments of insects^{1,2}. In both plants and animals they are transcribed in precise spatial patterns, often throughout development, and the particular combination of genes active in a cell specifies its identity. In several cases, expression late in development is both necessary and sufficient to specify identity^{3–5}. Because homeotic gene expression patterns are specified early in development as a result of complex genetic interactions (reviewed in refs 6, 7), this raises the problem of how these patterns are faithfully maintained and propagated through cell divisions during later development. Two alternative mechanisms can be envisaged. The first is that states of activity are maintained through cell-to-cell interactions, so that the identity of a cell influences that of its neighbour. This implies that identity is continually reassessed and is thus inconsistent with a state of determination. The second is that patterns are maintained in a cell-autonomous fashion, so that information as to the states of gene activity is inherited through cell division. This is consistent with a determined state, in which cells maintain fate irrespective of their environment. Regulation may proceed from the first to the second mechanism during development, as suggested for example by studies on the transcription of the *Drosophila* homeotic gene *engrailed*⁸.

In *Drosophila* embryogenesis, regional fate becomes determined early, at around the cellular blastoderm stage⁹, and alternative fates are selected and determined by the persistent activity of homeotic genes¹⁰. Studies of homeotic gene regulation have distinguished between classes of genes that control initial pattern specification and those involved in the maintenance of these patterns. Thus the initial patterns of expression of the homeotic genes in the Antennapedia (ANT-C) and Bithorax (BX-C) clusters are specified through interactions with gap and segmentation genes⁶. However, these early regulators are only expressed transiently and the maintenance of expression boundaries during later stages requires the activity of two antagonistic groups of genes, termed the Polycomb-group (Pc-G) and the trithorax-group (trx-G)^{11,12}. In Pc-G mutant embryos, patterns of homeotic gene expression are initially established correctly, but later in embryonic and adult development expression beyond the normal boundaries occurs^{13,14}. The wild-type Pc-G genes are therefore required not for pattern specification, but rather to ensure that repression is maintained in lineages where the target genes were initially inactivated. By contrast, the wild-type trx-G

products are required for maintenance of activity¹⁵. Transgenic studies using reporter genes indicate that the Pc-G and trx-G genes fix states of target gene activity such that they are heritable through cell division^{16,17}. These observations have led to the suggestion that the Pc-G and trx-G products are involved in imprinting a determined state, by fixing states of homeotic gene expression¹⁸.

In plants, experiments based on the analysis of marked cell lineages and on the effects of ablating cells suggest that in many cases the fate of plant cells is not determined, but rather is position-dependent and maintained through interactions between neighbouring cells¹⁹. However, a few studies suggest that around the time of floral induction, shoot meristems may become stably committed to form inflorescence or flowers^{20,21}. The role of homeotic genes is best understood with respect to flower development. Flowers typically contain four concentric whorls of organs with identity sepal (whorl 1), petal (whorl 2), stamen (whorl 3) and carpel (whorl 4). Based on the genetic and morphological analysis of floral homeotic mutants, a model has been proposed to account for the specification of organ identity in the different whorls based on combinatorial action of homeotic genes¹. In *Arabidopsis*, the *AGAMOUS* (AG) gene is required to specify stamen and carpel identity in whorls 3 and 4 respectively. Molecular isolation of the AG gene indicates that it encodes a protein belonging to the MADS box family of transcription factors, and that its RNA is confined to its domain of function in whorls 3 and 4 (refs 22, 23).

In this study we describe a novel mutation, *curly leaf* (*clf*), and show that the wild-type *CLF* gene is required to repress AG transcription in leaves, inflorescence stems and flowers. The deduced product of *CLF* has extensive structural homology with the product of a *Drosophila* Pc-G gene. In addition, *CLF* shows functional homology, because analysis of AG expression in *clf* flowers indicates that *CLF* is not required for the initial specification of the AG pattern, but rather to maintain repression during later stages of development. Together, these results suggest that Pc-G genes may act during flower development to promote fate determination.

Phenotype of *clf* mutants

A recessive mutation, designated the *clf-2* allele, was obtained from a transposon mutagenesis experiment²⁴. The mutation conferred pleiotropic effects on leaf and flower morphology and on flowering

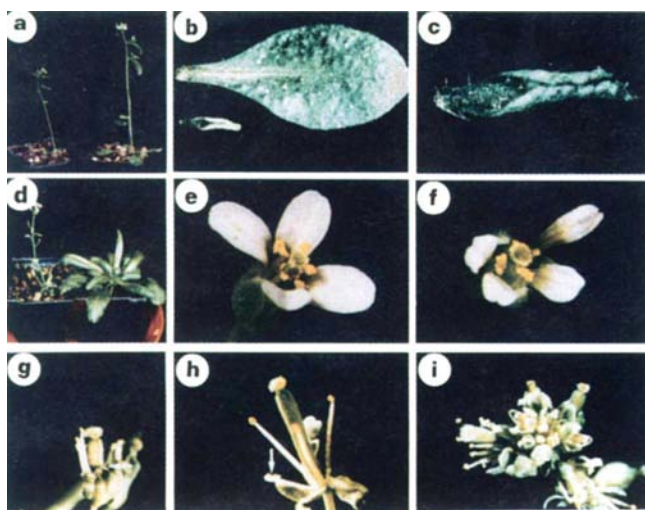


Figure 1 Comparison of *clf-2* mutant and wild-type *Arabidopsis* plants. **a**, Adult *clf-2* mutant (left) and wild-type plants growing under long-day conditions. The *clf-2* mutant is shorter and flowers earlier than wild type. **b**, Rosette leaf of wild-type (top) and *clf-2* mutant plant. The *clf-2* leaf is smaller and curled. **c**, Rosette leaf of *clf-2* mutant, exhibiting the curled leaf phenotype characteristic of the mutant. **d**, *clf-2* mutant (left) and wild-type (right) plants growing under short-day conditions. The *clf-2* mutant bolts (inflorescence stem elongates) and forms flowers earlier and after forming fewer leaves than the wild-type plant. **e**, A flower of a wild-type plant containing the normal arrangement of sepals, petals, stamens and carpels in whorls 1 to 4 respectively. **f**, A *clf-2* flower showing a weak mutant phenotype. The petals are narrower than those of wild type. **g**, A *clf-2* flower showing a strong mutant phenotype. The sepals are fused and have stigmatic papillae along their margins (arrow). Petal and stamen numbers are reduced. **h**, A *clf-2* flower showing a strong mutant phenotype. Petal number is reduced, a staminoid petal is indicated, with a patch of yellow anther-like tissue on the surface of the petal blade. **i**, *clf-2* inflorescence showing premature opening of flower bud.

time, but did not affect development of the roots, hypocotyl or cotyledons (Fig. 1).

The leaf was consistently the most severely affected organ. Whereas wild-type leaves had a broad, flat lamina (blade), that of *clf-2* plants was narrow and curled upwards along the longitudinal axis of the leaf (Fig. 1b, c). The extent of curling increased from base to apex of the plant, so that cauline leaves and leaves produced late in rosette development were more curled than the first leaves.

Wild-type plants flowered approximately four weeks later when grown under conditions of short-day illumination than under long days. By contrast, *clf-2* mutants flowered about three weeks earlier than wild type under short-day conditions (Fig. 1d), but were only two days earlier than wild-type under long-day conditions. There was therefore much less effect of day length on the flowering time of *clf-2* mutants than that of wild type. The stem internodes that elongate after flowering were shorter in *clf-2* than in wild-type plants, so that the height of the inflorescence was reduced (Fig. 1a).

The *clf-2* mutation also affected floral morphology. Unlike wild-type flowers, in which the sepals enclosed the floral bud until late in development, in *clf-2* mutants the sepals were frequently less curved and failed to enclose the floral buds so that the flowers were prematurely open (Fig. 1i). The petals of *clf* plants were also consistently smaller and narrower than in wild type (Fig. 1e, f). Mutant flowers also showed partial homeotic transformations of the sepals and petals towards carpels and stamens, respectively. Such transformations were more common in the petals than the sepals, and increased in frequency from the base to the apex of the inflorescence. The staminoid petals were intermediate between petals and stamens in shape, and had patches of yellow, anther-

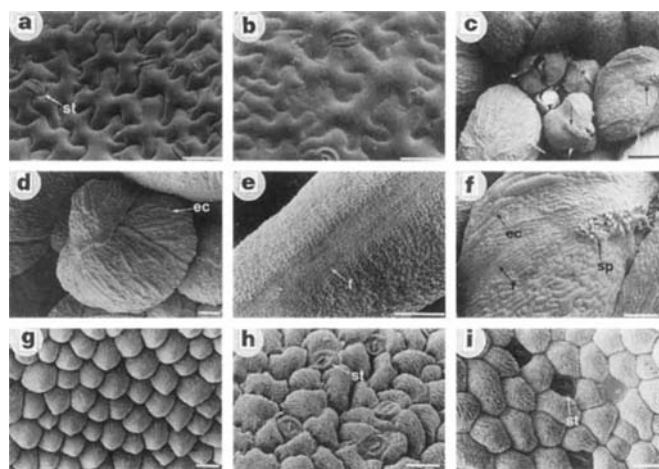


Figure 2 SEM analysis of epidermal cell types in *clf-2* and wild-type plants. **a**, Abaxial epidermis of a wild-type leaf. The epidermis consists of large, irregularly shaped cells interspersed with stomata (st). Scale bar, 30 μ m. **b**, A *clf-2* mutant leaf of similar age to leaf in **a**. The epidermal cells have similar morphology to those of wild type. Scale bar, 30 μ m. **c**, An inflorescence of a *clf-2* mutant. Two of the developing flowers (arrows) contain sepals fused at their margins with stigmatic papillae at their tips. Scale bar, 300 μ m. **d**, A wild-type flower bud. The sepals are not fused and contain characteristic highly elongated cells (ec). Scale bar, 100 μ m. **e**, Wild-type gynoecium. Files of small, regularly sized cells (f) were present at the margin of fusion of the two carpel valves. Scale bar, 100 μ m. **f**, Sepals of a *clf-2* mutant flower displaying the elongated cells (ec) characteristic of sepals, as well as the stigmatic papillae (sp) and files of small, regularly shaped cells (f) at the margin of fusion of the sepals that are characteristic of carpels. Scale bar, 100 μ m. **g**, Adaxial epidermal cells of a wild-type petal displaying their characteristic conical shape and the absence of stomata. Scale bar, 10 μ m. **h**, The Epidermis of a wild-type anther. The cells were flattened and had irregular outlines. Stomata (st) were present. Scale bar, 10 μ m. **i**, Adaxial epidermis of a staminoid *clf-2* mutant petal. The presence of stomata and the flattened shape of the cells resembled those of wild-type anthers. The patterns of cuticular thickening and the outline of the cells are intermediate between those of petals and stamens. Scale bar, 10 μ m.

like tissue on the white surface of the petal blade (Fig. 1h). In most severely affected flowers, second whorl organs were sometimes absent. The carpelloid sepals had several features that were characteristic of carpels, for example they were frequently fused along their margins, they had stigmatic papillae at their tips, and produced ovules on the margins (Fig. 1g). We did not observe any homeotic transformations in the third and fourth whorls of the flower. However, in severely affected flowers, there was a reduction in stamen number and the carpels were incompletely fused.

To analyse further the *clf-2* phenotype, the cell types present in the epidermis of the leaves and floral organs were analysed by scanning electron microscopy (SEM). The leaf epidermal cells of *clf-2* mutants were similar in morphology to those of wild type and did not show evidence of homeotic transformation (Fig. 2a, b). Analysis of the carpelloid sepals indicated partial homeotic transformation, because the epidermis contained cells typical of carpels as well as the very elongated narrow cells characteristic of sepals (Fig. 2c–f). The epidermal cells of the staminoid petals resembled those of stamens; whereas wild-type petals had conically shaped cells on their adaxial surface and lacked stomata, the *clf-2* staminoid petals had stomata and flattened cells similar to those of wild-type stamens (Fig. 2g–i).

The *clf-2* phenotype was sensitive to environment, in particular to day length and temperature. For example, under short days of 10-

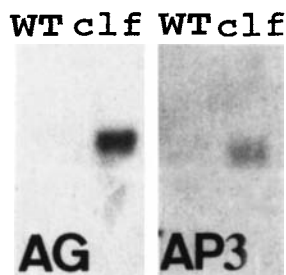


Figure 3 RNA gel blot analysis of *AG* and *AP3* expression. Total RNA was extracted from leaves of *clf-2* mutants and wild-type (WT) plants, blotted and probed with *AG* and *AP3* gene probes. *AG* and *AP3* mRNA is present in *clf-2* RNA, but not wild-type RNA.

hour photoperiods, flowering was delayed, the mutants were larger and showed weaker floral phenotypes than under long days. When grown at 16°C, *clf-2* plants flowered later than at 25°C and had leaves that were not curled and were only slightly smaller than wild type.

The *clf-2* phenotype resembled that described for accessions N313, N328 and N350 from the Nottingham *Arabidopsis* Stock Centre and complementation tests performed in the F1 and F2 generations confirmed that these lines carried recessive mutations allelic to *clf-2*, designated the *clf-1*, *clf-17* and *clf-19* alleles respectively. All of these alleles conferred a similar phenotype, with pleiotropic effects on leaf and floral morphology and on flowering time, and were similarly affected by environmental conditions. The *clf-2* mutation was as severe as any of the other three alleles, and was used for subsequent genetic analysis because it arose in the same genetic background (Landsberg *erecta*) as the other floral mutants. Detailed characterization of the other alleles will be presented elsewhere (P.P. *et al.*, manuscript in preparation).

CLF represses AG expression

The *clf* phenotype had striking similarity to that previously reported for transgenic plants in which the *AG* gene was constitutively expressed²⁵, suggesting that it might also result from misexpression of *AG*. Consistent with this, RNA blot analysis indicated that *AG* was expressed strongly in mature leaves of *clf-2* mutants but not in those of wild-type plants (Fig. 3). Similarly, weak misexpression of the floral homeotic gene *APETALA3* (*AP3*) was also detected in the leaves of *clf-2* mutants (Fig. 3). However, *AP3* expression is unlikely to contribute to the *clf* phenotype, because constitutive expression of *AP3* does not cause a phenotype in the leaves, and in the flowers causes stamen development in the fourth whorl, a phenotype very different to that of *clf* mutants²⁶. Misexpression of other floral homeotic genes, *LEAFY* (*LFY*), *APETALA1* (*API*) and *PISTILLATA* (*PI*), was not detected in the leaves of *clf-2* mutants in similar experiments (data not shown).

The pattern of *AG* expression was further localized by *in situ* hybridization of ³⁵S-labelled *AG* RNA probes to sections of mutant and wild-type plants. In longitudinal sections of vegetative (8-day-old) seedlings, no signal above background was detected in wild-type shoot apices and leaves (Fig. 4a, c). In similar sections of *clf-2* mutants, the shoot apical meristem did not express *AG* either. However, a weak signal was detected in the young leaf primordia that arose on the flanks of the meristem, and this became more intense in older primordia and in young leaves (Fig. 4b, d). In transverse sections of the first seedling leaves, *AG* expression was evenly distributed throughout the leaf, rather than confined to any specific cell layer or type (Fig. 4d). Unlike the leaves, there was little signal in cotyledons and hypocotyls other than an occasional weak one over the vasculature (Fig. 4b).

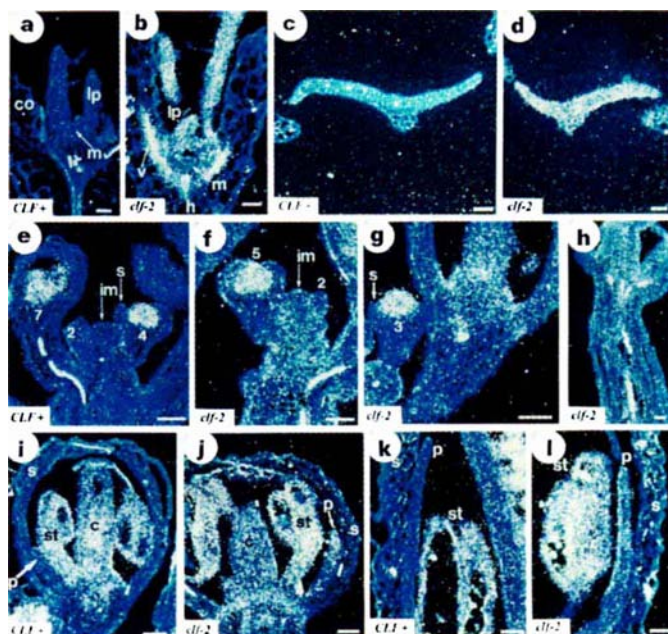


Figure 4 *In situ* hybridization with *AG* RNA probes. **a** and **b**, Longitudinal section of the apex of a wild-type (**a**) and *clf-2* mutant (**b**) seedling. *AG* mRNA is not present in the cotyledon (co), meristem (m) or leaf primordia (lp) of the wild-type seedling (**a**; scale bar, 40 µm). However, *AG* mRNA is present in the vasculature (v) of the cotyledon and hypocotyl (h), as well as in leaf primordia (lp) of the *clf-2* mutant, although it is absent from the meristem (m) (**b**; scale bar, 50 µm). **c** and **d**, Transverse section of leaves of wild-type (**c**) and *clf-2* plants. The wild-type leaf does not contain *AG* mRNA, but the transcript is evenly distributed throughout the *clf-2* mutant leaf. Scale bars, 80 and 70 µm, respectively. **e-g**, Longitudinal sections of inflorescences of wild-type plants and *clf-2* mutants. *AG* mRNA is not present in the inflorescence meristem (im) nor in stage-2 flowers of wild-type (**e**) or *clf-2* mutants (**f**). In stage-4 flowers of wild-type, *AG* mRNA is present in the centre of the flower primordium and is absent from the sepals (s in panel **e**). Similarly, in *clf-2* mutants, *AG* mRNA is present in the centre and absent from the sepals (s) of stage-3 (**g**) and stage-5 (**f**) flowers. Scale bars, 50 µm. **h**, Longitudinal section of the stem of a *clf-2* mutant. *AG* mRNA is present throughout the stem. Scale bar, 50 µm. **i-j**, Longitudinal sections of stage-9 flowers of wild-type (**i**) and *clf-2* mutant (**j**). *AG* mRNA is present in the stamens (st) and carpels (c) of wild-type and *clf-2* flowers. In *clf-2* mutants it is also present in petals (p; arrow in **j**) but is undetected in sepals (s), whereas in the wild type it is absent from both of these organs. Scale bars, 50 µm. **k, l**, Longitudinal sections of stage-12 flowers of wild-type (**k**) and *clf-2* mutant (**l**). Annotation is the same as for **i** and **j**. Scale bars, 50 and 40 µm, respectively.

The pattern of *AG* expression in wild-type inflorescences has been described previously²³. In brief, *AG* is not expressed in the inflorescence stem and meristem, nor in the young stage-1 and stage-2 floral meristems that arise on its flanks. Strong expression first occurs in the centre of stage-3 and -4 meristems but not in the emerging sepal primordia (Fig. 4e). During later stages of flower development, *AG* is persistently excluded from the sepal and petal primordia but is present in stamens and carpels in the centre of the flower.

In *clf-2* mutants, a similar pattern of *AG* expression was observed during the early stages of flower development. Thus, *AG* was not expressed in the inflorescence meristem nor in stage-1 and stage-2 floral meristems (Fig. 4f), and at stage 3, expression was confined to the centre of the meristem (Fig. 4g). During stages 5–8, *AG* continued to be excluded from sepals and from the emerging petal primordia. However, late in the development of the petals (stage 9 onwards), weak ectopic expression was detected (Fig. 4i–l).

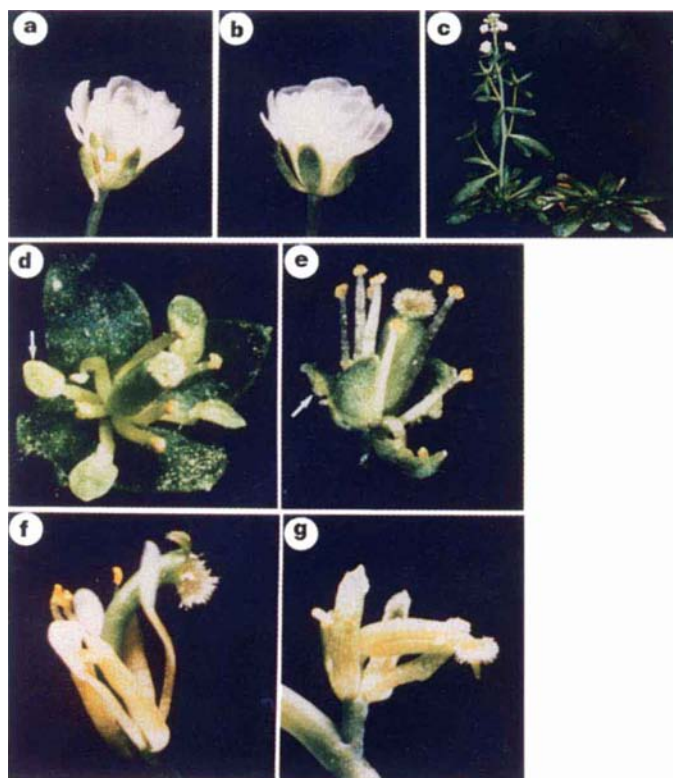


Figure 5 Genetic interactions between *CLF* and floral homeotic genes. **a**, An *ag-3* mutant flower. Petals rather than stamens are formed in the third whorl. In the fourth whorl, no carpels are present but further whorls of sepals and petals are formed. **b**, An *ag-3 clf-2* flower. The phenotype is identical to that of *ag-3*, demonstrating that *AG* is essential for the homeotic conversions present in the first two whorls of *clf-2* mutant flowers shown in Fig. 1. **c**, *ag-3 clf-2* (left) and *ag-3* (right) plants growing under short-day conditions. *ag-3 clf-2* mutants bolt and form flowers earlier than *ag-3* mutants. **d**, *ap2-1* mutant flower. Organs resembling cauline leaves form in the first whorl, and staminoid petals (arrow) develop in the second whorl. Organ identity in the third and fourth whorls is unaffected. **e**, *ap2-1 clf-2* mutant flower. The first whorl organs are more carpelloid than in *ap2-1*, containing stigmatic papillae along their margins and at their tips. The second whorl organs develop as stamens. **f**, *lug-8* mutant flower. The sepals and petals are narrow. The carpels are imperfectly fused. **g**, *lug-8 clf-2* mutant flower. Sepals are similar to those of the *lug* mutant. Petals are absent, and stamen number is reduced (to an average of two per flower).

In addition, *clf-2* plants showed *AG* expression in the inflorescence stem and in pedicels (Fig. 4h).

Taken together, these observations indicated that the wild-type *CLF* gene is required to repress *AG* in hypocotyls, cotyledons, leaves, inflorescence stems and petals. In addition to *CLF*, the *APETALA2* (*AP2*) and *LEUNIG* (*LUG*) genes are also required to repress *AG* in flowers^{23,27}. However, our analysis of *AG* transcription in *clf-2* mutants indicates that *CLF* acts later in flower development than *AP2* or *LUG*, consistent with the weak phenotype of *clf* flowers compared to those of *ap2* or *lug* mutants. In strong *ap2* mutants, *AG* RNA appears in mid-stage-2 flowers, and is present in all floral whorls in stage-3 and older flowers, and in *lug* mutants a similar pattern is observed, but expression in whorls 1 and 2 is weaker^{23,27}. This pattern of ectopic *AG* expression differs from that detected in *clf* mutants, because the early pattern of *AG* transcription in *clf* inflorescence apices and floral meristems is the same as in wild type, but subsequently *AG* is expressed in the inflorescence stem and in petals late in development (stage 9 onwards). These results suggest that *CLF* is not required for the initial specification of the *AG*

expression pattern, but rather for its maintenance during later development. We did not detect *AG* expression in the first whorl of *clf* flowers, although the occasional carpelloid of the sepals and the observation that *clf-2 ag* double mutants have normal sepals (see below) suggest that *AG* activity is present in the first whorl. Our failure to detect *AG* mRNA in these organs might be explained if it was present at very low abundance, or only for a limited period of developmental time.

Genetic interactions between *ag* and *clf*

The *clf* mutant phenotype was very similar to that conferred by constitutive expression of *AG* in transgenic plants, raising the possibility that the *clf* phenotype was entirely a consequence of ectopic *AG* expression. In this case, *ag* mutations would be predicted to be epistatic to *clf*. To test this, we examined plants that were doubly mutant for *clf-2* and the *ag-3* allele, which confers a severe loss-of-function *ag* phenotype. The flowers of the double mutant had normal sepals and petals and were indistinguishable from those of *ag-3* single mutants (Fig. 5a, b). Therefore, in flowers *CLF* is required only to repress *AG* activity and, unlike *AP2* and *API*, it does not have an independent homeotic function to specify sepal and petal identity. When grown under long days, the double mutants were also very similar to *ag-3* single mutants with respect to leaf morphology, size and flowering time. In addition, the *clf* phenotype was sensitive to *AG* dosage; plants homozygous for *clf-2* and heterozygous for *ag-1* usually showed only weak leaf curling (data not shown). Taken together, these results suggested that the *clf* phenotype was dependent upon *AG* activity.

However, under short-day conditions which delayed flowering, some differences were apparent. Thus the double mutants were slightly earlier flowering than *ag-3* single mutants (Fig. 5c), they were smaller and had leaves which showed extremely weak curling. Similar results were obtained in double mutants of *clf-2* and *ag-1*, another severe *ag* allele (data not shown). One possible explanation for the incomplete epistasis of *ag-1* and *ag-3* to *clf-2* is that, in addition to *AG*, *CLF* regulates other genes that have weak effects on leaf curling and flowering time. It is unlikely that these include any of the floral homeotic genes *LFY*, *API*, *LUG*, *AP2*, *AP3* or *PI*, because double-mutant combinations of each of these with *clf-2* did not influence the severity of the *clf* leaf or flowering time phenotype. Furthermore, *LFY*, *API* and *PI* are not ectopically expressed in the leaves of *clf* mutants, and constitutive expression of *AP3* does not affect leaf morphology or flowering time²⁶. An alternative explanation for the weak phenotype of the double mutants is that *ag-1* and *ag-3* are not completely null alleles and their products retain a weak activity when expressed in vegetative tissues. Consistent with this, the lesions at these alleles are point mutations predicted to affect splicing of the fourth *AG* intron, so that the amino-terminal region of the *AG* protein, which includes the MADS box DNA-binding domain, is likely to be unaffected²⁰ (M. Yanofsky, personal communication). In the absence of definitive null *ag* alleles, we were unable to exclude this possibility.

CLF encodes a Pc-G protein

The *clf-2* mutation was identified during a transposon mutagenesis experiment in which derivatives of the maize transposons *Activator* (*Ac*) and *Dissociation* (*Ds*) were mobilized in transgenic *Arabidopsis* lines²⁴. A single copy of *Ds*, at a different position from that in the wild-type progenitor, was present in *clf-2* mutants. Two lines of evidence suggested that the *clf-2* mutation was caused by this *Ds* insertion. First, the *Ds* transposon and the *clf-2* mutation co-segregated and we were unable to detect recombination between them²⁴. Second, the mutation was unstable in the presence of *Ac*, which is a characteristic feature of *Ds*-induced alleles. This instability was recognized because plants homozygous for *clf-2* and carrying *Ac* produced some phenotypically wild-type progeny (revertants), whereas this never occurred with *clf-2* homozygotes that did not

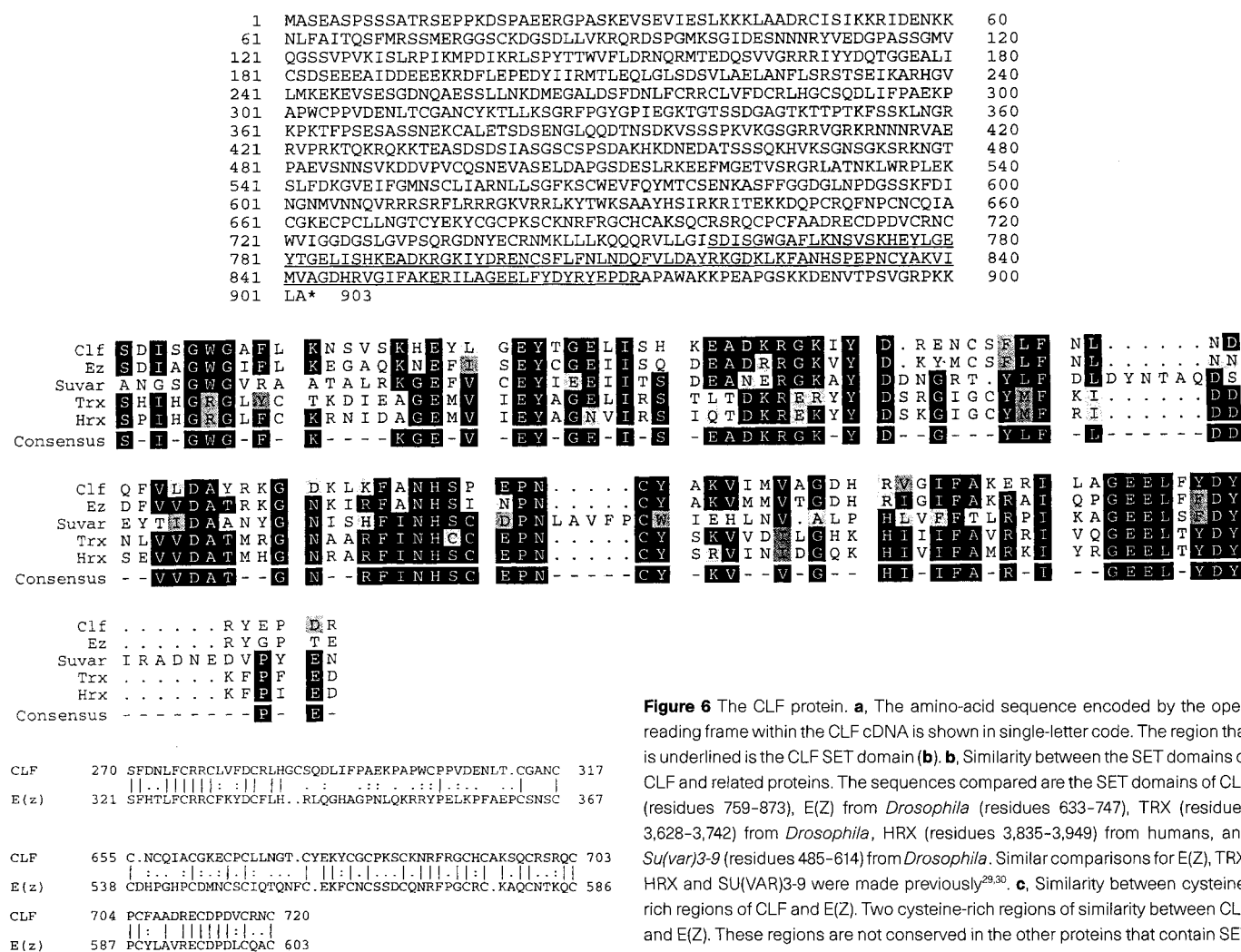


Figure 6 The CLF protein. **a**, The amino-acid sequence encoded by the open reading frame within the CLF cDNA is shown in single-letter code. The region that is underlined is the CLF SET domain (**b**). **b**, Similarity between the SET domains of CLF and related proteins. The sequences compared are the SET domains of CLF (residues 759–873), E(Z) from *Drosophila* (residues 633–747), TRX (residues 3,628–3,742) from *Drosophila*, HRX (residues 3,835–3,949) from humans, and Su(var)3-9 (residues 485–614) from *Drosophila*. Similar comparisons for E(Z), TRX, HRX and SU(VAR)3-9 were made previously^{29,30}. **c**, Similarity between cysteine-rich regions of CLF and E(Z). Two cysteine-rich regions of similarity between CLF and E(Z). These regions are not conserved in the other proteins that contain SET domains and are shown in **c**.

carry Ac. We therefore isolated clones of genomic DNA that flanked the *Ds* insertion and presumably spanned the *CLF* locus (see Methods). To confirm that this was derived from the *CLF* locus, a DNA fragment spanning the *Ds* insertion site was hybridized to genomic DNA blots of DNA extracted from wild type, *clf-2* mutants and 10 independently isolated revertant plants. In mutants, the wild-type fragment was absent and fragments of novel size were detected. All 10 revertant plants showed the same pattern as the wild type. *Ds* often excises imprecisely, causing sequence alterations (footprints)²⁸. The point of insertion of *Ds* in *clf-2* was sequenced in five of the revertants. Of the five sequences, three were identical to wild type, whereas one contained an insertion of 3 base pairs (bp) and the other a deletion of 21 bp. The absolute correlation between phenotypic reversion and excision of the *Ds* transposon confirmed that *clf-2* was caused by insertion of the transposon.

To identify the *CLF* transcript, complementary DNA libraries prepared from young flowers were screened with a genomic fragment spanning the *Ds* insertion at *clf-2* and several clones were isolated. The longest clone contained a 3.05-kilobase (kb) insert and detected a single, weak transcript of about 3.2 kb when hybridized to RNA blots containing poly(A)⁺ RNA from 8-day-old seedlings (data not shown). This is consistent with the cDNA representing the full-length transcript, taking into account a poly(A) tail of 100–200 bases. Comparison of the cDNA and genomic sequences indicated that the transcription unit contained 18 exons and that the *Ds* insertion at *clf-2* was located in the first intron, three

nucleotides downstream of the 5' exon–intron junction. The presence of the transposon within the transcription unit confirmed that the cDNA derived from the *CLF* transcript.

The cDNA sequence contained a long open reading frame (ORF) starting at position +97, with an ATG flanked by sequences that conformed to the consensus for initiation of translation in plants²⁹. The ORF encoded a potential protein, CLF, of 902 amino acids (Fig. 6a). Comparison of the amino-acid sequence of CLF to protein sequence databases using the BLAST program³⁰ revealed extensive homology between CLF and the product of the *Drosophila* gene *Enhancer of zeste* (*E(z)*), a member of the Polycomb-group of genes. There were three regions that were conserved between the two proteins. First, the carboxy terminus of CLF contained a 115-amino-acid region, the SET domain, previously recognized as a conserved region in the products of three *Drosophila* genes, *E(Z)*, *TRITHORAX* (*TRX*) and *SU(VAR)3-9*, all involved in transcriptional regulation^{31,32}. The SET domains of CLF and *E(Z)* are more similar to each other (65% identity) than either is to the SET domains of *TRX* or *SU(VAR)3-9* (Fig. 6b). Second, residues 655–720 of CLF showed 46% amino-acid identity with a region of *E(Z)* (residues 538–603) that was previously identified as being rich in cysteine residues, but with an arrangement unlike that of zinc-fingers or any other family of cysteine clusters (Fig. 6c). The spacing of the cysteines was extremely well conserved in CLF. Third, residues 270–317 of CLF contained seven cysteines with a similar spacing to a region of *E(Z)* (residues 321–367) containing five cysteines, and

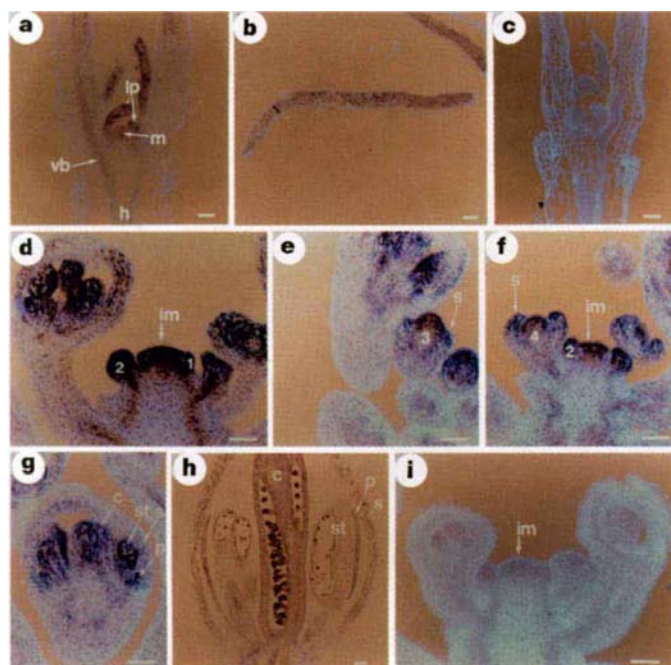


Figure 7 RNA *in situ* hybridization with *CLF*. Longitudinal sections of wild-type Landsberg *erecta* plants probed with antisense (a, b, c, d, e, f, g, h) and sense (c and i) *CLF* probes. a, An eight-day-old seedling. *CLF* mRNA is present in leaf primordia (lp), meristem (m) and the vasculature (vb) of cotyledons and hypocotyl (h). b, The first leaf of a seedling. *CLF* mRNA is evenly distributed throughout the leaf. c, No hybridization is detected when a seedling section is probed with a *CLF* sense probe. d, e and f, Inflorescences of Landsberg *erecta*. *CLF* mRNA is present in the inflorescence meristem (im) as well as throughout stage-1 and stage-2 flowers (d and f). Flowers of stages 3 and 4 contain *CLF* mRNA in the central region and in the sepals. g, h, Flowers of stages 8 and 12, respectively. *CLF* mRNA is present in carpels (c), stamens (st), petals (p) and sepals (s). i, No hybridization is detected when an inflorescence section is probed with a *CLF* sense probe. Scale bar is 50 μ m in each panel.

there was a small region of sequence similarity at the N termini of these regions (Fig. 6c). These two cysteine-rich regions were unique to *CLF*, *E(Z)* and mammalian *E(Z)* homologues, and were not found in other SET domain proteins in the databases. Sequence analysis of temperature-sensitive *E(z)* alleles suggests that both cysteine-rich regions are functionally significant³³. The *CLF* protein also contained a basic domain (residues 610–626), which conformed to a bipartite nuclear localization signal (NLS) and was in a position corresponding to a potential NLS in the *E(Z)* protein³¹.

CLF* is insufficient to repress *AG

To investigate the pattern of *CLF* expression, sections of wild-type plants were hybridized *in situ* with labelled *CLF* RNA probes (Fig. 7). No signal was detected in seedlings or inflorescences probed with the sense strand of *CLF* (Fig. 7c, i). When *CLF* antisense probes were hybridized to longitudinal sections of 8-day-old seedlings, a strong signal was detected throughout the apical meristem, leaf primordia and leaves (Fig. 7a, b). A weak signal occurred in the vasculature of the hypocotyl. In sections of inflorescences, the signal observed with the *CLF* probe was weaker than those obtained with probes for the floral homeotic genes *AG* or *LFY* (data not shown), suggesting that *CLF* was expressed at a comparatively low level. A strong signal was present in the inflorescence meristem, whereas there was weaker expression in the stem in the provasculature (Fig. 7d). *CLF* was strongly expressed throughout the young stage-1 and -2 floral meristems that arose on the flanks of the apex. In stage-3 and -4

flowers, *CLF* was expressed in the emerging sepal primordia as well as in the dome of the floral meristem (Fig. 7e, f). During stages 6–9, *CLF* was expressed strongly in developing petal, stamen and carpel primordia, whereas expression in the sepals was weaker (Fig. 7g). Late in floral development, at stage 12, *CLF* continued to be expressed weakly in all floral whorls, with strongest expression in petals and in ovules (Fig. 7h). Expression of *CLF*, at least at the transcriptional level, is therefore insufficient to repress *AG*, because *CLF* RNA is present in all four whorls throughout floral development and overlaps with *AG* expression in the third and fourth whorls. Specificity of *CLF* action must therefore involve interaction with other genes, for which likely candidates include the *AP2* and *LUG* genes, which also repress *AG* in floral whorls 1 and 2.

That *AP2* and *CLF* act in the same pathway is suggested by the observation that the double mutant of *clf-2* and the weak *ap2-1* allele had flowers with a more severe phenotype than either parent, but a weaker phenotype than the strong alleles *ap2-9* and *ap2-2* (Fig. 5d, e). The strongest enhancement was in the second whorl, which in *ap2-1* flowers contained organs intermediate between petals and stamens but in the double mutant consisted of stamens. In the first whorl, organs were leaf-like or sepal-like but had more carpeloid features than in *ap2-1*. This phenotype increased in severity from base to apex of the inflorescence. The strong *ap2-2* or *ap2-9* alleles were epistatic to *clf-2* in the flowers. These results suggest that *AP2* and *CLF* act in a common pathway to repress *AG*. Furthermore the level of *AP2* is poised at a critical level in the *clf* background, as recessive *ap2* mutations become semidominant, consistent with the genes acting in the same pathway, or a direct interaction between the two proteins.

Flowers of *lug* mutants have sepals that are transformed towards stamens and carpels, and petals that are either staminoid or absent²⁷. In *clf-2 lug-8* flowers, second-whorl organs were completely absent, representing a weak enhancement of the *lug* phenotype (Fig. 5f, g). In other floral whorls, organs resembled those of *lug* single mutants. The double mutant therefore showed an additive phenotype, suggesting that *LUG* acts independently of *CLF*.

Role of Pc-G proteins in plants

The *CLF* and *E(Z)* proteins show extensive structural homology. The conserved SET domain and cysteine-rich regions do not have similarity with DNA-binding domains or other motifs of known function, and their biochemical function is unclear. However, the three *Drosophila* proteins, SU(VAR)3-9, *E(Z)* and TRX, which contain SET domains, are all involved in the heritable maintenance of transcriptional states³², and there is indirect evidence that they do so by modifying chromatin structure^{34–37}. The presence of a SET domain in three proteins that are implicated in complexes that affect chromatin structure suggests that *CLF* might also act on chromatin, for example packaging the *AG* target so that transcriptional activators can not gain access or are unable to interact with general transcription initiation factors. Furthermore, methylation of the *AG* gene might be required for its repression by *CLF*, because although transgenic plants in which expression of DNA methyl transferase is reduced show a variety of phenotypes^{38,39}, some of them are similar to *clf* mutants and this is also correlated with ectopic *AG* expression³⁸. The SET domain of *CLF* also shows homology with a gene that was identified as an *Arabidopsis* expressed-sequence tag (Genbank H36381), indicating that there are other plant genes containing SET domains, although their function is as yet unknown.

In addition to structural similarity, *CLF* and *E(Z)* also display functional similarity in the repression of homeotic gene expression. Like other Pc-G members, *E(z)* is necessary for maintenance of transcriptional repression of many genes, including homeotic genes in the ANT-C and BX-C complexes. For example, in embryos that lack *E(z)*⁺ activity, the *Ultrabithorax* (*Ubx*) gene is expressed in its normal pattern until 5–7 hours after egg-laying but subsequently

maintenance of the anterior boundary breaks down and it is eventually expressed throughout the germ band¹⁴. The wild-type *E(z)* gene is therefore required not for the initial specification of the *Ubx* pattern, but rather to ensure that repression is maintained during later development. Similarly, in *clf* mutant flowers, AG is only expressed ectopically late in development (Fig. 4), suggesting that *CLF* is also involved in maintaining repression rather than establishing the transcriptional pattern of its target. Thus, the *AP2* and *LUG* genes are required for the initial specification of the AG transcription pattern from stage 2 in floral meristem development, whereas *CLF* is required to maintain repression late in development from stage 9 onwards. Unlike *Drosophila* embryogenesis, in which pattern specification is achieved by factors expressed transiently, in *Arabidopsis* flowers *AP2* RNA is expressed throughout development⁴⁰. However, studies using the temperature-sensitive *ap2-1* allele indicate that *AP2* activity is not required beyond the early stages of flower development³. It is therefore possible that *CLF* provides a memory of AG repression during later stages.

It is unclear whether *CLF* has a similar maintenance function in leaves, because AG is not expressed vegetatively in wild-type nor *ap2* mutant plants, and no other genes have yet been shown to be necessary to prevent vegetative AG transcription. However, the *AP2* gene is a member of a small gene family⁴⁰, at least one other member of which is also expressed in leaves, so it is possible that these genes have a redundant role in repressing AG in vegetative meristems or leaf primordia and that *CLF* acts to maintain this. Alternatively, *CLF* might function directly to specify AG repression in leaf primordia.

Several studies indicate that the Pc-G genes are widely conserved, and that in vertebrates they are also required for stable repression of homeotic genes^{41,42}. Because the homeotic gene clusters of insects and vertebrates are considered to be homologous⁴³, the use of Pc-G genes as repressors presumably occurred in a common ancestor and has been conserved between the two lineages. The similarities in *CLF* and *E(z)* action suggest that Pc-G gene function may also be conserved in plants. This is striking, because the homeotic target genes in animals and plants are structurally unrelated and encode either homeobox or MADS box class transcription factors, respectively. The use of Pc-G genes as repressors of homeotic genes may therefore have evolved independently in plants, presumably in response to the common problem of ensuring that patterns specified early in development are faithfully propagated through subsequent cell divisions.

The maintenance of homeotic gene repression by a Pc-G gene provides one mechanism for determination of the fate of floral cells, although it is unclear how widely this occurs in plant development. For example, laser ablation of meristematic cells in roots alters the fate of neighbouring cells, suggesting that at least in early development fate is not determined in roots⁴⁴. Unlike the *Drosophila* Pc-G genes, which regulate many homeotic and other target genes¹⁴, *CLF* appears largely to be a regulator of AG. The expression of other floral homeotic genes may therefore be maintained by alternative mechanisms. For example, maintenance of high-level expression of the floral homeotic genes that specify petal and stamen identity involves positive autoregulation^{26,45}. Alternatively, determination of cell fate may involve regulation of homeotic genes by additional Pc-G and trx-G genes in plants. □

Methods

Growth conditions. Plants were grown on a 2:1 mixture of Levington's M3 compost:grit in greenhouses. During winter, supplementary light was provided so that the minimum day length was 16 h. For measurement of flowering time, plants were grown at 20 °C in Gallenkamp controlled-environment cabinets with short-day (SD) and long-day (LD) conditions as described⁴⁶. For the floral organ counts, plants were grown under continuous illumination at 25 °C on a 1:1:1 mixture of vermiculite:perlite:soil²⁷.

Genetic analysis. All mutations were present in the Landsberg *erecta* genetic background. The *clf-2* allele was obtained from a transposon mutagenesis

experiment²⁴. Stocks carrying the *ag-1*, *ap1-1*, *ap2-1* and *pi-1* mutations were obtained from the Nottingham *Arabidopsis* Stock Centre; all other mutant stocks were supplied by members of our laboratories. To construct double mutants, *clf-2/+* heterozygotes were crossed as males to females homozygous for floral mutations, except in the case of sterile *ag* mutants when *ag/+* heterozygotes were used. Double mutants segregated in the expected 1:15 ratio in the F₂ progeny. To confirm the double-mutant phenotypes, seed was collected from individual F₂ *clf* plants, and in 2/3 of the resulting F₃ families the double mutant segregated 1:3. The effects of AG or *AP2* dosage were confirmed by progeny tests of presumed heterozygotes.

Scanning electron microscopy. Tissues were fixed in glutaraldehyde, dehydrated through an ethanol series and critical-point dried as described³. Specimens were sputter-coated with gold and examined in a Camscan MR4 scanning electron microscope using an accelerating voltage of 16 kV. In some cases (Fig. 2a–d, f), fresh tissues were frozen in nitrogen slush at –190 °C, ice crystals were removed by sublimation at –90 °C; specimens were then sputter-coated and examined on a cold stage fitted to the microscope.

DNA and RNA blot analysis. DNA extraction and DNA-blot analysis were done by established methods²⁴. Total RNA was purified as described⁴⁷, and poly(A)⁺ RNA was isolated using the polyAtract system (Promega). Molecular probes were isolated from the clones pCIT 565 for AG, pD793 for *AP3* (ref. 48) and pCDJ4 for *CLF*.

Cloning and sequencing of *CLF*. Genomic DNA flanking the *Ds* insertion in the *clf-2* allele was amplified by inverse polymerase chain reaction (IPCR) and cloned into pKR vector²⁴. DNA adjacent to the 5' end of *Ds* was amplified using the oligonucleotide primers B34 (5'-ACGGTCGGTACGGGATTTCCCAT-3') and D74 (GGATATACAAAACGGTAAACGGAAACG), followed by a second round of amplification using the nested primers D73 (ref. 24) and E4 (ref. 24). For DNA adjacent to the 3' end of *Ds*, the primer sets were B39 (TTTCGTTTCCGTCGCCGAAGTTAAATA) and C133 (ATCTCCACTGACGT AAGGGATGACG), then C133 and D71 (CCGTTACCGACCGTTTTCAT CCCTA). The IPCR products were used as molecular probes to screen λ libraries prepared from genomic DNA of either Ler or Columbia ecotype (gifts from M. Anderson and C. Chang, respectively), and clones λP1 and λGJ12 spanning the *CLF* locus were obtained. These clones were further characterized by DNA-blot analysis to obtain a detailed map of the *CLF* locus. To isolate cDNA clones, a λZAPII (Stratagene) library prepared from RNA of young flowers⁴⁹ was screened with probes derived from *CLF* genomic clones. 12 clones were obtained from a screen of about 12 × 10⁶ PFU, the longest clone was sequenced in full and several other clones were partially sequenced to confirm that they derived from the same transcription unit.

In situ hybridization. AG antisense probes were prepared from a pCIT 565 template as described²³. For *CLF*, a 0.9-kb *Xba*I/*Xho*I fragment of the cDNA clone pCDJ4 (nucleotides 778–1,675) was subcloned into the plasmids Bluescript SK+ and KS+ (Stratagene) to produce clones pQUIT1 and pQUIT2, respectively. Sense and antisense probes were synthesized using T7 RNA polymerase and pQUIT2 linearized with *Xba*I or pQUIT1 linearized with *Xho*I, respectively. For *in situ* hybridization with S³⁵-labelled probes, previously described procedures²³ were followed except that tissue fixation was limited to 1 h. For experiments using digoxigenin-labelled probes, the methods used for tissue fixation, probe synthesis, hybridization and detection were as described⁵⁰. In both cases, tissue sections were stained by immersing in 0.1% calcofluor for 5 min, rinsing in water for 5 min and mounting in Entellan (BDH Merck).

Received 3 October 1996; accepted 28 January 1997.

- Coen, E. S. & Meyerowitz, E. M. The war of the whorls: genetic interactions controlling flower development. *Nature* **353**, 31–37 (1991).
- Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**, 565–570 (1978).
- Bowman, J. L., Smyth, D. R. & Meyerowitz, E. M. Genes directing flower development in *Arabidopsis*. *Plant Cell* **1**, 37–52 (1989).
- Carpenter, R. & Coen, E. S. Floral homeotic mutations produced by transposon-mutagenesis in *Antirrhinum majus*. *Genes Dev.* **4**, 1483–1493 (1990).
- Struhl, G. Genes controlling segmentation specification in the *Drosophila* thorax. *Proc. Natl Acad. Sci. USA* **79**, 7380–7384 (1982).
- Ingham, P. W. The molecular genetics of embryonic pattern formation in *Drosophila*. *Nature* **335**, 25–33 (1988).
- Weigel, D. & Meyerowitz, E. M. The ABCs of floral homeotic genes. *Cell* **78**, 203–209 (1994).
- Heemskerk, J., DiNardo, S., Kostriken, R. & O'Farrell, P. H. Multiple modes of engrailed regulation in the progression towards cell fate determination. *Nature* **352**, 404–410 (1991).
- Simcox, A. A. & Sang, J. H. When does determination occur in *Drosophila* embryos? *Dev. Biol.* **97**, 212–221 (1983).
- Lawrence, P. A. & Morata, G. In *Insect Development* (ed. Lawrence, P. A.) 132–148 (Blackwell, Oxford, 1976).

11. Simon, J. Locking in stable states of gene expression: transcriptional control during *Drosophila* development. *Curr. Opin. Cell Biol.* **7**, 376–385 (1995).
12. Kennison, J. A. Transcriptional activation of *Drosophila* homeotic genes from distant regulatory elements. *Trends Genet.* **9**, 75–79 (1993).
13. Struhl, G. & Akam, M. Altered distributions of *Ultrabithorax* transcripts in *extra sex combs* mutant embryos of *Drosophila*. *EMBO J.* **4**, 3259–3264 (1985).
14. Jones, R. S. & Gelbart, W. M. Genetic analysis of the *Enhancer of zeste* locus and its role in gene regulation in *Drosophila melanogaster*. *Genetics* **126**, 185–199 (1990).
15. Breene, T. R. & Harte, P. J. *trithorax* regulates multiple homeotic genes in the bithorax and Antennapedia complexes and exerts different tissue-specific, parasegment-specific and promoter-specific effects on each. *Development* **117**, 119–134 (1993).
16. Grindhart, J. G. J. & Kaufman, T. C. Identification of *polycomb* and *trithorax* group responsive elements in the regulatory region of the *Drosophila* homeotic gene *sex combs reduced*. *Genetics* **139**, 797–814 (1995).
17. Muller, J. Transcriptional silencing by the Polycomb protein in *Drosophila* embryos. *EMBO J.* **14**, 1209–1220 (1995).
18. Paro, R. Imprinting a determined state into the chromatin of *Drosophila*. *Trends Genet.* **6**, 416–421 (1990).
19. Weigel, D. & Doerner, P. Cell–cell interactions—taking cues for the neighbors. *Curr. Biol.* **6**, 10–12 (1996).
20. Irish, E. E. & Nelson, T. M. Identification of multiple stages in the conversion of maize meristems from vegetative to floral development. *Development* **112**, 891–898 (1991).
21. Bradley, D., Vincent, C., Carpenter, R. & Coen, E. Pathways for inflorescence and floral induction in *Antirrhinum*. *Development* **122**, 1535–1544 (1996).
22. Yanofsky, M. F. *et al.* The protein encoded by the *Arabidopsis* homeotic gene *AGAMOUS* resembles transcription factors. *Nature* **346**, 35–39 (1990).
23. Drews, G. N., Bowman, J. L. & Meyerowitz, E. M. Negative regulation of the *Arabidopsis* homeotic gene *AGAMOUS* by the *APETALA2* product. *Cell* **65**, 991–1002 (1991).
24. Long, D. *et al.* The maize transposable element system *Ac/Ds* as a mutagen in *Arabidopsis*: Identification of an *albino* mutation induced by *Ds* insertion. *Proc. Natl Acad. Sci. USA* **90**, 10370–10374 (1993).
25. Mizukami, Y. & Ma, H. Ectopic expression of the floral homeotic gene *AGAMOUS* in transgenic *Arabidopsis* plants alters floral organ identity. *Cell* **71**, 119–131 (1992).
26. Jack, T., Fox, G. L. & Meyerowitz, E. M. *Arabidopsis* homeotic gene *APETALA3* ectopic expression—transcriptional and posttranscriptional regulation determined floral organ identity. *Cell* **76**, 703–716 (1994).
27. Liu, Z. & Meyerowitz, E. M. *LEUNIG* regulates *AGAMOUS* expression in *Arabidopsis* flowers. *Development* **121**, 975–991 (1995).
28. Sutton, W. D., Gerlach, W. L., Schwartz, D. & Peacock, W. J. Molecular analysis of *ds* controlling element mutations at the *ADH1* locus of maize. *Science* **223**, 1265–1268 (1984).
29. Lüttke, H. A. *et al.* Selection of AUG initiation codons differs in plants and animals. *EMBO J.* **6**, 43–48 (1987).
30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, B. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
31. Jones, R. S. & Gelbart, W. M. The *Drosophila* Polycomb-group gene *Enhancer of zeste* contains region with a sequence similarity to *trithorax*. *Mol. Cell. Biol.* **13**, 6357–6366 (1993).
32. Tschiersch, B. *et al.* The protein encoded by the *Drosophila* position-effect variegation suppressor gene *Su(var)3–9* combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J.* **13**, 3822–3831 (1994).
33. Carrington, E. A. & Jones, R. S. The *Drosophila* *Enhancer of zeste* gene encodes a chromosomal protein: examination of wild-type and mutant protein distribution. *Development* **122**, 4073–4083 (1996).
34. Paro, R. & Hogness, D. S. The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of *Drosophila*. *Proc. Natl Acad. Sci. USA* **88**, 263–267 (1991).
35. Rastelli, L., Chan, C. S. & Pirota, V. Related chromosome binding sites for *zeste*, suppressors of *zeste* and *Polycomb* group proteins in *Drosophila* and their dependence on *Enhancer of zeste* function. *EMBO J.* **12**, 1513–1522 (1993).
36. Tsukiyama, T., Becker, P. B. & Wu, C. ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. *Nature* **367**, 525–532 (1994).
37. Côté, J., Quinn, J., Workman, J. L. & Peterson, C. L. Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science* **265**, 53–60 (1994).
38. Finnegan, E. J., Peacock, W. J. & Dennis, E. S. Reduced DNA methylation in *Arabidopsis thaliana* results in abnormal plant development. *Proc. Natl Acad. Sci. USA* **93**, 8449–8454 (1996).
39. Ronemus, M. J., Galbiati, M., Ticknor, C., Chen, J. C. & Dellaporta, S. L. Demethylation-induced developmental pleiotropy in *Arabidopsis*. *Science* **273**, 654–657 (1996).
40. Jofuku, D., van den Boer, B., Van Montagu, M. & Okamoto, J. Control of *Arabidopsis* flower and seed development by the homeotic gene *APETALA2*. *Plant Cell* **6**, 1211–1225 (1994).
41. van der Lugt, N. M. T. *et al.* Posterior transformation, neurological abnormalities, and severe hematopoietic defects in mice with a targeted deletion of the *bmi-1* protooncogene. *Genes Dev.* **8**, 757–769 (1994).
42. Akaseka, T. *et al.* A role for *mel-18*, a Polycomb group-related vertebrate gene, during the anteroposterior specification of the axial skeleton. *Development* **122**, 1513–1522 (1996).
43. Krumlauf, R. *Hox* genes in vertebrate development. *Cell* **78**, 191–201 (1994).
44. van den Berg, C., Willemsen, V., Hage, W., Weisbeek, P. & Scheres, B. Determination of cell fate in the *Arabidopsis* root meristem by directional signalling. *Nature* **378**, 62–65 (1995).
45. Tröbner, W. *et al.* *Globosa*—A homeotic gene which interacts with *deficiens* in the control of *Antirrhinum* floral organogenesis. *EMBO J.* **11**, 4693–4704 (1992).
46. Putterill, J., Robson, E., Lee, K., Simon, R. & Coupland, G. The *CONSTANS* gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. *Cell* **80**, 847–857 (1995).
47. Coen, E. S., Carpenter, R. & Martin, C. Transposable elements generate novel spatial patterns of gene expression in *Antirrhinum majus*. *Cell* **47**, 285–296 (1986).
48. Jack, T., Brockman, L. L. & Meyerowitz, E. M. The homeotic gene *APETALA3* of *Arabidopsis thaliana* encodes a MADS box and is expressed in petals and stamens. *Cell* **68**, 683–697 (1992).
49. Weigel, D., Alvarez, J., Smyth, D. R., Yanofsky, M. F. & Meyerowitz, E. M. *LEAFY* controls floral meristem identity in *Arabidopsis*. *Cell* **69**, 843–859 (1992).
50. Coen, E. S. *et al.* *Floricula*: a homeotic gene required for flower development in *Antirrhinum majus*. *Cell* **63**, 1311–1322 (1990).

Acknowledgements. We thank K. Findlay for Fig. 2d, and R. Jones, E. Coen and our colleagues for discussion. We are grateful to N. Hartley, K. Patel and G. Murphy for help with the DNA sequence analysis. J.G. was funded by a PMB fellowship from the BBSRC for collaborative projects between groups in the UK and USA. P.P. was funded by the Government of Thailand, and M.M. by the Spanish Government. The work in the laboratory of E.M.M. was funded by the US NIH and in the laboratory of G.C. by a PMB2 grant from the BBSRC.

Correspondence and requests for materials should be addressed to G.C. (e-mail: couplnd@bbsrc.ac.uk). The *CURLY LEAF* cDNA sequence has been submitted to EMBL under accession number Y10580.

A candidate dust disk surrounding the binary stellar system BD+31°643

Paul Kalas* & David Jewitt

Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822, USA

Planetesimals are thought to form within the dense dust disks that surround young stars¹, although this process has not been observed directly. In contrast, the presence of dust disks around older, main-sequence stars can be used to infer that planetesimal formation has indeed occurred; dust is rapidly depleted in circumstellar environments, so a continuous supply of dust by the sublimation and collisional erosion of pre-existing planetesimals^{2,3} is required to maintain a disk over the lifetime of the central star. The composition of the dust in such disks should provide clues regarding the composition of the planetesimals, and the structure of the disks (for example, gaps in the dust distribution) could be linked to physical interactions with unseen planets. To date, only one main-sequence star— β Pictoris—has been shown to have a dust disk that can be resolved optically⁴. Here we report the optical image of a candidate dust disk surrounding a main-sequence binary stellar system, BD+31°643. If the existence of this dust disk is confirmed by future observations, it would imply that binary stars may possess stable environments for planetesimal formation.

To establish criteria for identifying circumstellar nebulosities as candidates for near-edge-on disks, we use β Pic as a standard. The β Pic disk is a nearly linear, optical reflection nebulosity bisected by the star (Fig. 1), with a radius of at least 1,000 AU (60 arcsec)⁵, but with significant asymmetries in the spatial distribution of dust^{6,7}. Other explanations for the nebulosity, such as outflow phenomena, have been ruled out⁸. We therefore adopt the following criteria for finding disk candidates: (1) linear nebulosity bisected by the star, (2) morphological symmetry between the two sides of the nebulosity, (3) exclusion of outflow phenomena, and (4) exclusion of chance alignment with background objects.

As part of a current optical search for circumstellar disks⁹, we obtained coronagraphic images of the B5V binary star BD+31°643. Located in the Perseus II star-forming region ~ 330 pc from the Sun, it illuminates the 2–3 arcmin scale reflection nebulosity IC348. IC348 was studied photographically as early as 1922¹⁰, and more recently with a charge-coupled device (CCD)¹¹ and near-infrared array¹², but never with a stellar coronagraph.

Observations were taken on 18 December 1995 UT at the University of Hawaii 2.2-m telescope on Mauna Kea. A coronagraph was used to artificially eclipse BD+31°643, as well as to suppress diffracted light within the telescope. We used broadband V and R Mould colour filters centred at 545 and 647 nm, respectively (approximating the Kron–Cousins colour system). Images were recorded with a $2,048 \times 2,048$ pixel CCD at a plate scale of 0.41 arcsec per pixel. Our image resolution was ~ 1.4 arcsec. In addition to BD+31°643, several stars devoid of circumstellar nebulosity were coronagraphically imaged under nearly identical conditions to sample the point-spread function (PSF).

The coronagraphic images of BD+31°643 reveal a nebulosity within 20 arcsec of the star that is linear, bisected by the star, and $1\text{--}2$ mag arcsec⁻² brighter than the arcminute-scale, diffuse nebulosity of IC348 (Figs 1, 2a). Subsequent, near-infrared images (taken without a coronagraph) resolve BD+31°643 into a previously

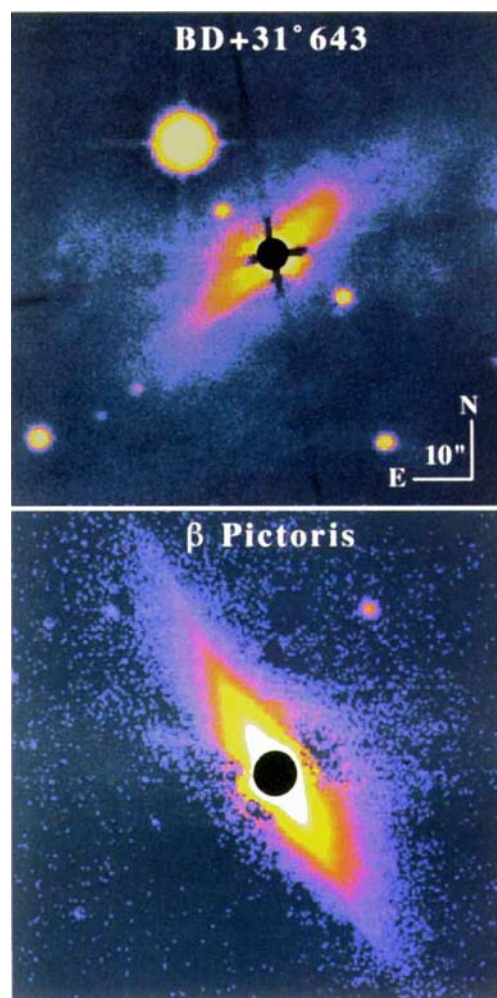


Figure 1 False-colour, R-band images of BD+31°643 (top) and β Pic (bottom; ref. 6) with effective integration times of 210 s and 890 s, respectively. A coronagraph was used to artificially eclipse BD+31°643 and β Pic with 3-arcsec and 6.5-arcsec diameter opaque spots, respectively (in these images, software-enlarged masks are shown, and the wires supporting the spots have been digitally removed). To study the disk near the edges of the spots, appropriately scaled and registered PSFs were subtracted from the central stars, as well as from the bright star (Gingrich 12) to the northeast of BD+31°643. For BD+31°643, an artificial PSF was created by fitting a seventh-order polynomial to the radial profile of a nearby, dust-free comparison star. We investigated possible subtraction errors due to the binarity of the BD+31°643 PSF and found them to be negligible beyond 4 arcsec radius. For Gingrich 12, its own light profile was sampled to the northeast (away from BD+31°643) to generate a similar artificial PSF for subtraction of scattered light. The net effect is a significant reduction of scattered light near BD+31°643 which allows the detection of the disk-like structure down to the edge of the occulting mask. The midplane of the BD+31°643 disk intersects the central binary at position angle $PA = 131^\circ \pm 5^\circ$. Near-infrared data at K' ($\lambda = 2.2 \mu\text{m}$) obtained by K. Jim (University of Hawaii) resolve the two stellar components. They have roughly equal magnitudes ($\Delta K' = 0.16 \pm 0.05$ mag), lie along a line with $PA = 16.8^\circ \pm 2.0^\circ$ and are separated by 0.60 ± 0.02 arcsec.

known¹³ binary. Isophotes of the nebulosity show an approximately symmetric, disk-like morphology (Figs 1, 2c), though the southeast extension appears longer and narrower than the northwest extension by $\sim 15\%$. The midplane surface brightness gradient is approximately symmetric between the two extensions from 4 to 20 arcsec projected radius (Fig. 2b). The midplane gradient gradually flattens at smaller projected radii and can be characterized by two least-squares,

* Present address: Max-Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany.

Table 1 Physical properties

	BD+31°643*	β Pic†
Spectral type	B5V	A5V
Stellar mass	$2 \times 5 M_{\odot}$	$1.5 M_{\odot}$
Apparent visual magnitude	+8.5	+3.8
Absolute visual magnitude	-1.2	+2.7
Effective temperature	15,400 K	8,200 K
Luminosity	$2 \times 830 L_{\odot}$	$6 L_{\odot}$
Age	$\leq 10^4$ yr	$\geq 10^4$ yr
Distance	330 pc	16.4 pc
V surface brightness at 6"	$20.0 \text{ mag arcsec}^{-2}$	$16.3 \text{ mag arcsec}^{-2}$
Colour relative to star	Blue	Neutral
Midplane surface brightness	$r^{-0.3}$ (4"–7"), $r^{-1.5}$ (7"–20")	$r^{-3.8}$ – $r^{-4.1}$ (6"–16")
Radial number density	$r^{-1.7}$ (4"–7"), $r^{-0.3}$ (7"–20")	$r^{-2.8}$ – $r^{-3.5}$ (6"–24")
Observed radius	20" (6,600 AU)	60" (1,000 AU)
Length asymmetry	15%	20%
Inclination	0°–10°	0°–5°
Depleted region (radius)	2,300 AU	1–30 AU
Dust cross-section	$8 \times 10^{31} \text{ cm}^2$ (5"–20")	$7 \times 10^{26} \text{ cm}^2$ (5"–40")

In rows 2 and 6, we assume that the two components of BD+31°643 have approximately equal mass and luminosity. In rows 11, 12 and 17, numbers in parentheses indicate the range of disk radii in arcsec where measurements were made and/or properties derived.

* Parameters obtained from refs 16, 28–30.

† Parameters obtained from refs 3, 6–22.

power-law fits from projected radii 4 to 7 arcsec, and 7 to 20 arcsec (Table 1).

This morphology is inconsistent with collimated bipolar outflows (such outflows typically have an asymmetric appearance at visible wavelengths and occur during the pre-main-sequence phase of stellar evolution)^{14,15}, and there is no spectroscopic evidence for outflow in previous studies of this star¹⁶. That BD+31°643 has been established as the primary source of illumination for the larger-scale IC348 nebosity, and that 12- μ m and 60- μ m emission from this region is centred on the binary¹⁶, both indicate that the features nearest the binary are due to dust physically associated with BD+31°643 and not a chance alignment with a background object. The bright nebosity meets our criteria for the optical detection of a near-edge-on circumstellar disk candidate.

We use a model of scattered light from dust arranged with the geometry of a disk¹⁷ to fit the observed morphology and midplane surface brightness. Isophote shapes are fitted by model disks having inclinations $\leq 10^\circ$ from edge-on (Fig. 2c). For example, an isotropically scattering model dust disk, inclined 10° away from edge-on, with an $r^{+1.7}$ increase in number density from 4 to 7 arcsec (r is radial distance), and an $r^{-0.7}$ decrease in number density from 7 to 20 arcsec fits our data (Fig. 2b, c). The radial number density gradient beyond 7 arcsec radius is flatter than derived for the β Pic disk ($\sim r^{-3}$; ref. 6), but comparable to our Solar System's zodiacal dust cloud (r^{-1} ; ref. 18).

The model fits imply a centrally depleted region that is significantly larger in BD+31°643 than in β Pic (Table 1)⁷. Tidal interaction with the binary may create the central depletion of material near BD+31°643, as well as the asymmetries in disk structure¹⁹. The binarity of BD+31°643 has been recorded since 1881, but only a fraction of the orbit has been observed¹³. Further observations are required to clarify the orbital parameters, as well as to confirm the existence of the central depletion. However, we note that our model fits do not provide a unique explanation for the flattening of the midplane surface brightness gradient. Spatial variations in particle properties, such as the size distribution and composition, are not incorporated into the models, but may have a significant effect on the appearance of the nebosity.

Table 1 compares the physical properties of the two stars and nebulosities. BD+31°643 is 20 times more distant than β Pic, 3.9 magnitudes (absolute) brighter, but is viewed through 2.8 magnitudes of visual extinction²⁰. We measure $V - R = 0.2 \pm 0.2$ for both

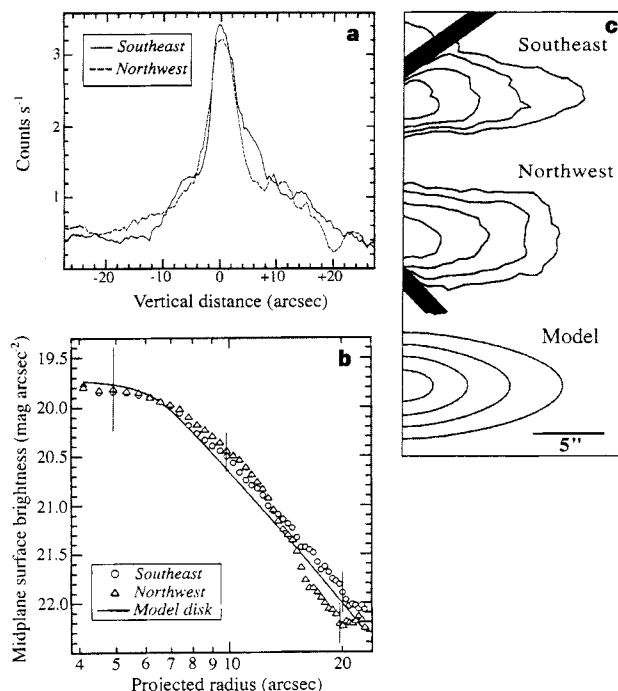


Figure 2 V-band data and model fits to the disk-like nebosity around BD+31°643. **a**, Profiles perpendicular to the disk midplane at $r = 8$ arcsec (projected radius). The disk appears as a sharp feature ($|z| \leq 6$ arcsec, where z is the vertical distance from the disk midplane), superimposed on a broad bump ($6'' < |z| \leq 15''$) and the diffuse, arcminute-scale IC348 nebosity ($|z| > 15''$). We define the outer radius of the disk, $r = 20$ arcsec, as the point where the disk can no longer be distinguished from the broad bump. **b**, Surface brightness along a strip 1.2-arcsec wide centred along the midplanes of each extension. To isolate disk light, the background nebosity was sampled 6 arcsec north and south of the midplane (see **a**), the mean background was fitted with a third-order polynomial, which was then used to subtract light along the direction of disk midplane. The downturn in surface brightness within ~ 7 arcsec radius is seen in both extensions and in our R-band data (it is also evident before the subtraction of the background nebosity). The scatter of points along the curves indicates the magnitude of statistical uncertainty ($\sim 0.1 \text{ mag arcsec}^{-2}$). However, the systematic uncertainty due to the background nebosity is the dominant source of error, which we estimate and show as error bars at radii of 5, 10 and 20 arcsec. We also plot the midplane surface brightness of a model circumstellar disk which is depleted of material within 7 arcsec radius (see text). **c**, Surface brightness isophotes in V beginning from 5 arcsec projected radius (contour interval, $0.5 \text{ mag arcsec}^{-2}$; outer contour, $21.5 \text{ mag arcsec}^{-2}$). The data were rotated and transposed such that the star lies to the left, and the north side of each extension is up. Dark bars represent regions blocked by the occulting spot wires. The subtraction of background nebosity introduces an uncertainty in the morphology of the outer contour of $\sim 15\%$. The isotropic scattering model disk discussed in the text is shown for comparison.

the disk and ambient nebosity at 10 arcsec radius, whereas the star has $V - R = 0.5$ (refs 11, 21). The bluer colour of the nebosity is consistent with previous findings¹¹ and indicates that the dominant particle size, a , is smaller than the wavelength of light, that is, $a \approx 0.1 \mu\text{m}$. However, the uncertainty of our measurements precludes an accurate assessment of possible colour gradients due to changes in particle size or composition.

We estimate from our optical data that the total dust scattering cross-section of BD+31°643 is three orders of magnitude larger than β Pic (Table 1). However, as BD+31°643 occupies a larger volume, the radial optical depths along the midplanes of both disks are $\tau \approx 10^{-3}$. If we assume dominant grain sizes of $a \approx 0.1 \mu\text{m}$ for BD+31°643, and $a \approx 1 \mu\text{m}$ for β Pic²², then we infer from the total

scattering cross-sections approximate dust masses of 10^{-6} solar masses ($10^{-6} M_{\odot}$) for BD+31°643, and $10^{-8} M_{\odot}$ for β Pic (assuming a particle density $\rho = 1 \text{ g cm}^{-3}$ and an albedo $A = 0.1$). These must be lower limits because we have not measured scattered light within 5 arcsec radius of either star, nor have we considered the likely presence of larger, unseen bodies.

In the absence of a significant gaseous component to the disk, forces such as Poynting–Robertson (PR) drag and radiation pressure will deplete the small dust surrounding BD+31°643 on short timescales. For a particle radius $a = 0.1 \mu\text{m}$, density $\rho = 1 \text{ g cm}^{-3}$, distance from the star $r = 2,000 \text{ AU}$ and luminosity $L_{\star} = 1,660 L_{\odot}$ (where $L_{\odot}$ is the solar luminosity), the PR decay time is $t_{\text{PR}} = 0.2 \times 10^6 \text{ yr}$. Even more significant is radiation pressure, which will blow out dust in an optically thin system if the ratio of radiation pressure to central gravity is $\beta > 1$. For a given particle size and composition, β scales very roughly as the ratio of stellar luminosity to stellar mass and is 40 times larger for BD+31°643 than for β Pic. Extrapolating from the β values calculated for various grain sizes and compositions around β Pic²³, we estimate that $\sim 0.1\text{-}\mu\text{m}$ grains at 2,000 AU from BD+31°643 will be blown out to 6,000 AU on 10^3-yr timescales. The dominance of the radiation force over stellar gravity implies that the observed, disk-like structure is not centripetally supported and is therefore fundamentally different from the 10–100 AU-scale accretion disks detected around pre-main-sequence stars²⁴.

The age of BD+31°643 is probably greater than the $6 \times 10^5 \text{ yr}$ required for a BV5 star to reach the main sequence²⁵, but less than the $(5\text{--}7) \times 10^6 \text{ yr}$ age of the IC348 cluster²⁶. The fact that the lifetime of $0.1\text{-}\mu\text{m}$ dust is orders of magnitude shorter than the age of the star argues that the dust within 6,000 AU is replenished. Erosion of larger bodies has been suggested as a source of replenishment for the β Pic disk⁸; adopting this suggestion, the inclination distribution of dust observed around BD+31°643 may reflect the inclination distribution of unseen parent objects. These parent bodies have either been partially cleared, or were prevented from forming near the stars due to tidal interactions, PR drag, collisions or other destructive forces. They are present at $\sim 10^3 \text{ AU}$ radii where they continue to replenish $0.1\text{-}\mu\text{m}$ -sized dust grains. If we assume an age of $\sim 10^6 \text{ yr}$ and a dust lifetime of $\sim 10^3 \text{ yr}$, then the disk has been replenished 1,000 times over. Using the observed disk mass of $\sim 10^{-6} M_{\odot}$, we estimate that the mass of precursor objects is $\sim 10^{-3} M_{\odot}$ (a lower limit). The parental dust cloud (that is, the larger-scale IC348 nebulosity) still surrounds the disk and binary, and the entire system may represent an intermediate stage of evolution between embedded protostars (for example, the protobinary system IRAS16293–2422²⁷) and isolated, main-sequence stars such as β Pic.

Additional observations are required to test the disk interpretation of the nebulosity surrounding BD+31°643, to measure the gas content associated with the dust, and to clarify the relationship of the disk-like structure with the surrounding dust cloud. If its disk nature is confirmed, BD+31°643 represents an exceptional opportunity to examine the evolution and stability of planetesimals located around a binary stellar system. □

Received 20 May 1996; accepted 10 January 1997.

- Weidenschilling, S. J. & Cuzzi, J. N. *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 1031–1060 (Univ. Arizona Press, Tucson, 1993).
- Nakano, T. *Mon. Not. R. Astron. Soc.* **230**, 551–571 (1988).
- Backman, D. E. & Paresce, F. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 1253–1304 (Univ. Arizona Press, Tucson, 1993).
- Smith, B. A. & Terrile, R. J. *Science* **226**, 1421–1424 (1984).
- Smith, B. A. & Terrile, R. J. *Bull. Am. Astron. Soc.* **19**, 829 (1987).
- Kalas, P. & Jewitt, D. *Astron. J.* **110**, 794–804 (1995).
- Lagage, P. O. & Pantin, E. *Nature* **369**, 628–630 (1994).
- Artymowicz, P. in *Proc. IAP Meeting on Circumstellar Dust Disks and Planet Formation* (eds Ferlet, R. & Vidal-Madjar, A.) 47–65 (Editions Frontieres, Paris, 1994).
- Kalas, P. & Jewitt, D. *Bull. Am. Astron. Soc.* **25**, 1353 (1993).
- Hubble, E. *Astrophys. J.* **56**, 400–438 (1922).
- Witt, A. N. & Schild, R. E. *Astrophys. J. Suppl. Ser.* **62**, 839–852 (1986).
- Zinnecker, H., McCaughrean, M. J. & Wilking, B. A. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 429–495 (Univ. Arizona Press, Tucson, 1993).

13. Eggen, O. J. *Astron. J.* **68**, 483–514 (1963).
14. Edwards, S., Ray, T. & Mundt, R. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 567–602 (Univ. Arizona Press, Tucson, 1993).
15. Staudte, H. J. & Elsasser, H. *Astron. Astrophys. Rev.* **5**, 165–238 (1993).
16. Snow, T. P., Hanson, M. M., Seab, C. G. & Saken, J. M. *Astrophys. J.* **420**, 632–642 (1994).
17. Kalas, P. & Jewitt, D. *Astron. J.* **111**, 1347–1355 (1996).
18. Lamy, P. L. & Perrin, J.-M. *Astron. Astrophys.* **163**, 269–286 (1986).
19. Artymowicz, P. & Lubow, S. H. *Astrophys. J.* **421**, 651–667 (1994).
20. Strom, S. E., Strom, K. M. & Carrasco, L. *Publ. Astron. Soc. Pacif.* **86**, 798–805 (1974).
21. Fernie, J. D. *Publ. Astron. Soc. Pacif.* **95**, 782–785 (1983).
22. Paresce, F. & Burrows, C. *Astrophys. J.* **319**, L23–L25 (1987).
23. Artymowicz, P. *Astrophys. J.* **335**, L79–L82 (1988).
24. Basri, G. & Bertout, C. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 543–566 (Univ. Arizona Press, Tucson, 1993).
25. Iben, I. *Astrophys. J.* **141**, 993–1014 (1965).
26. Lada, E. A. & Lada, C. J. *Astron. J.* **109**, 1682–1696 (1995).
27. Walker, C. K., Carlstrom, J. E. & Biegling, J. H. *Astrophys. J.* **402**, 655–666 (1993).
28. Underhill, A. & Doazon, V. *B Stars With and Without Emission Lines* (SP-456, NASA, Washington DC, 1982).
29. Schmidt-Kaler, Th. *Landolt-Bornstein, New Series, Group VI* (Springer, Berlin, 1982).
30. Racine, R. *Astron. J.* **73**, 233–245 (1968).

Acknowledgements. We thank G. Herbig and D. Backman for comments. The observations were made possible with the assistance of K. Jim, J. Grady, B. Zuckerman and E. Becklin. P.K. was supported by the NASA Graduate Student Researchers Program; D.J. was funded by the NASA Origins of Solar Systems program.

Correspondence should be addressed to P.K.

High-energy ions produced in explosions of superheated atomic clusters

T. Ditmire, J. W. G. Tisch, E. Springate, M. B. Mason, N. Hay, R. A. Smith, J. Marangos & M. H. R. Hutchinson

Blackett Laboratory, Prince Consort Road, Imperial College of Science, Technology and Medicine, London SW7 2BZ, UK

Efficient conversion of electromagnetic energy to particle energy is of fundamental importance in many areas of physics. A promising avenue for producing matter with unprecedented energy densities is by heating atomic clusters, an intermediate form of matter between molecules and solids¹, with high-intensity, ultra-short light pulses^{2–4}. Studies of noble-gas clusters heated with high-intensity ($>10^{16} \text{ W cm}^{-2}$) laser pulses indicate that a highly ionized, very high temperature micro-plasma is produced. The explosion of these superheated clusters ejects ions with substantial kinetic energy^{3–5}. Here we report the direct measurement of the ion energy distributions resulting from these explosions. We find, in the case of laser-heated xenon clusters, that such explosions produce xenon ions with kinetic energies up to 1 MeV. This energy is four orders of magnitude higher than that achieved in the Coulomb explosion of small molecules⁶, indicating a fundamental difference in the nature of intense laser–matter interactions between molecules and clusters. Moreover, it demonstrates that access to an extremely high temperature state of matter is now possible with small-scale lasers.

Such experiments have become possible owing to recently developed high peak power, femtosecond lasers which are based on chirped pulse amplification and are capable of producing focused light intensity up to $10^{14}\text{--}10^{19} \text{ W cm}^{-2}$ (ref. 7). These high intensities have been used to study the production of highly charged ions from multi-photon ionization of individual atoms⁸, and the optical ionization of small (2–10 atom) molecules, in which the resulting Coulomb explosion produces ions with kinetic energy of up to $\sim 10\text{--}100 \text{ eV}$ (ref. 9). At the same time, the production of hot ($\sim 1\text{-keV}$), high-density plasmas by the intense irradiation of a solid has also been the subject of detailed studies¹⁰. Though the nature of the interaction of intense light with single atoms and bulk solid targets has been the subject of extensive investigation over the past decade,

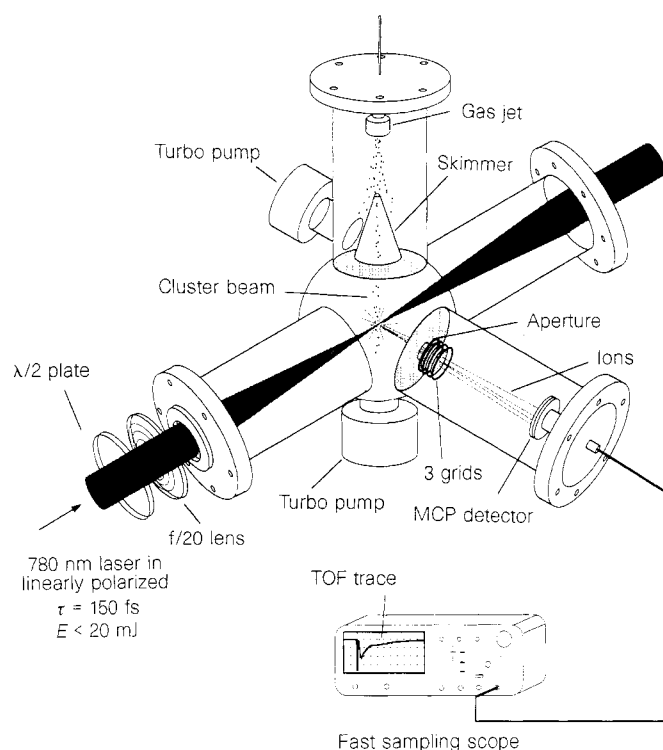


Figure 1 Experimental configuration for measuring ion energies of exploding clusters. The energies of the ions produced by the interaction of the laser pulse with the clusters were determined by the flight time of the ions in a field-free drift tube. The charge-state distribution of the ions as a function of ion kinetic energy was measured by varying the voltage on the middle grid of three closely spaced grids (~ 3 -mm spacing) placed along the drift tube (TOF, time-of-flight; MCP, microchannel plate detector; E , laser energy; τ , laser pulse width).

only recently have the dynamics of intense laser interactions with the intermediate state of matter, clusters, come under study.

Preliminary experiments on high-intensity laser-cluster interactions have suggested that this interaction is much more energetic than laser/isolated-atom interactions^{2-4,11,12}, and may be even more energetic than laser-solid interactions¹³. Recent experiments examining the X-ray emission from plasmas formed by the irradiation of noble-gas (Ar, Kr and Xe) clusters have indicated that, because of the enhancement of laser-driven collisional excitation within the cluster, efficient production of X-rays is possible^{2,3,11}. High-energy Coulomb explosion of small clusters ionized by femtosecond pulses has also been observed¹². The ion energies measured in these experiments ranged from 500 to 1,000 eV for ions with charges as high as 8+.

In an intensely irradiated cluster, optically and collisionally ionized electrons undergo rapid collisional heating for the short time (< 1 ps) before the cluster disassembles in the laser field⁵. Our recent studies⁴ of the electron energy spectra produced by the high-intensity irradiation of large Xe clusters with a 150-fs laser pulse indicated that collisional heating within the cluster can produce electrons with energy of up to 3 keV, an energy much higher than that typical of solid-target plasmas¹⁴. A sharp peak in the measured electron energy spectrum suggested that the cluster micro-plasma exhibited a resonance in the heating by the laser pulse similar to the giant resonance seen in metallic clusters¹⁵. A small amount of cluster expansion during the laser pulse lowers the electron density to bring the near-infrared laser light into resonance with the free-electron plasma frequency in the cluster. This resonance greatly increases the laser electric field density within the cluster, and the laser absorption rate is enhanced, rapidly heating the electrons on a very fast

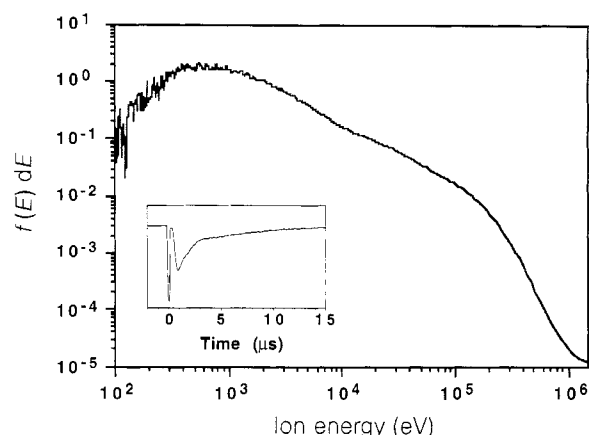


Figure 2 Ion energy spectrum from 65-Å ($\sim 2,500$ -atom) Xe clusters irradiated by a peak intensity of $2 \times 10^{16} \text{ W cm}^{-2}$, derived from the ion TOF trace shown as an inset. (The first peak in the TOF trace corresponds to fast (> 2 -keV) electrons; the feature at later times corresponds to the ions. Note that electrons with energy < 2 keV do not reach the MCP detector as the front plate is charged to $-2,000$ V).

(< 10 -fs) timescale to a highly non-equilibrium superheated state (with mean energies of many keV). Charge separation of the hot electrons inevitably leads to a very fast expansion of the cluster ions. This process is fundamentally different from low-intensity photo-fragmentation of clusters¹⁶ and far more energetic than the Coulomb explosion of small molecules⁹.

To search for this ion explosion, our experiment (Fig. 1) focused a high-power, 150-fs pulse from a Ti:sapphire laser¹⁷ into a time-of-flight chamber to a peak intensity of up to $2 \times 10^{16} \text{ W cm}^{-2}$ where it intercepted a low-density beam of Xe clusters. The time-of-flight trace directly yields an energy distribution function for the ions. In Fig. 2, we show the energy spectrum of Xe ions produced from the irradiation of 65-Å Xe clusters. This energy spectrum is radically different from previously observed ion energy spectra from the intense-field fragmentation of molecules or small clusters. Most surprising is the production of significant numbers of very hot ions with energies up to 1 MeV. This energy is four orders of magnitude higher than has previously been observed in the Coulomb explosion of molecules⁶ and about 1,000 times higher than the energy of Ar ions ejected in the Coulomb explosion of small clusters irradiated at lower intensity¹². The average ion energy of the distribution shown in Fig. 2, defined as $\langle E \rangle = \int E f(E) dE / \int f(E) dE$, is 45 ± 5 keV. Thus the average laser energy deposited in the cluster per ion is substantial. Furthermore, we find that the ions are ejected with equal energy in all directions with respect to the laser polarization, an artefact of the spherical geometry of the exploding cluster⁵. Again, this is in stark contrast to the Coulomb explosion of small linear molecules¹⁸.

By varying the gas-jet backing pressure we controlled the average size of clusters with which the laser interacts. In general, we find that both the maximum and the average ion kinetic energy increase with increasing cluster size, though both vary slowly with changing cluster size. Even in Xe clusters with as few as ~ 400 atoms (diameter of ~ 35 Å) the maximum observed ion energy is 250 keV and the average ion energy is 28 keV. However, at lower backing pressure, below ~ 1 bar, we observe no hot ions. Rayleigh-scattering measurements⁴ in the jet indicate that large clusters (> 100 atoms) form only with backing pressure > 1 bar. This transition to hot-ion production with large clusters points to an important change in the dynamics of the cluster expansion between small and large clusters. The slow decrease in ion energy with decreasing cluster size is related to the fact that small clusters disassemble faster in the laser pulse⁵. As a consequence, the electrons are heated to a lesser degree

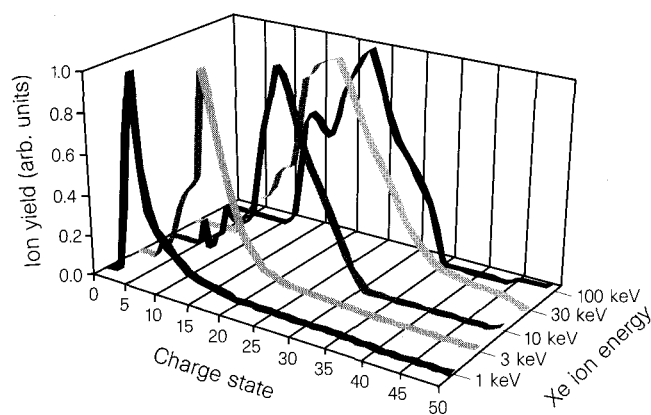


Figure 3 Measured charge-state distribution of 65-Å Xe clusters irradiated by a peak laser intensity of $2 \times 10^{18} \text{ W cm}^{-2}$ for ion kinetic energies of 1, 3, 10, 30 and 100 keV.

by the laser, and the cluster micro-plasma explodes with a lower velocity.

Another striking feature of the ions produced in the laser-cluster interaction is the ionization to very high charge states (a phenomenon previously seen in experiments^{2,3,12,19}). Measured charge-state distributions of Xe ions are shown in Fig. 3 for five different ion energies. In general the average charge state of the ions increases as the ion energy increases. For the high-energy ions, the peak of the charge-state distribution is around a charge state of 18+ to 20+ though a significant number of ions with charge state between 30+ and 40+ are observed. These are charge states far above the maximum that can be produced by simple field ionization of the Xe atoms⁸. In the cluster the high-temperature electrons created through laser-driven heating rapidly strip the ions by collisional ionization.

The observed charge states and energies of the ions are very much like those observed in the hot-electron-driven expansion of a laser-heated, solid target plasma^{20,21}. These high ion energies are the result of fast electron charge separation and rapid hydrodynamic expansion of the plasma into vacuum²². Previous studies of laser/solid-target interaction with intensities similar to those used in our experiment indicate that the characteristic ion energies from high-Z targets are²³ 100–200 keV. The close similarity between our data and the hot-ion data of high-energy, solid-target experiments indicates that, for clusters even as small as a few hundred atoms, the dynamics of the expansion appear to deviate from molecular behaviour and approach that of solid-target, hot-electron-driven expansions.

The efficiency with which the absorbed laser energy is converted to ion energy in the cluster explosion is remarkable. Though the majority of the laser energy is deposited into the electrons within the cluster during the laser pulse, the charge separation of these hot electrons accelerates the ions radially, and, as a result, after the explosion of the cluster most (>90%) of the laser energy deposited in the cluster is transferred to the ions. Both electrons and ions ultimately reach a velocity given by the sound speed of the cluster plasma, $c_s = \sqrt{ZkT_e/m_i}$ (where kT_e is the electron thermal energy and m_i is the ion mass)²², but most of the resulting kinetic energy is contained in the ions owing to their much greater mass. We expect that $\frac{1}{2}m_{Xe}c_s^2$ will be $\sim 50 \text{ keV}$ (for $kT_e \approx 2.5 \text{ eV}$ (ref. 4) and Xe ions with mass m_{Xe} and $Z \approx 20+$), in good agreement with our observed average Xe ion energies.

Our ion energy and charge-state measurements indicate that in a gaseous medium of clusters, the average energy deposited per atom may be as high as many tens of keV. Thus, in a gas of clusters with a modest average atomic density of $\sim 1 \text{ atm}$, a density typical near the output of most cluster-forming gas jets²⁴, virtually all of the laser

energy from a short pulse laser could be deposited in the gas within a focal volume, a prediction confirmed by recent laser energy absorption measurements in clustering gas jets²⁵. But unlike solid targets (in which much of the deposited laser energy is rapidly conducted into the cold, bulk solid^{10,26}), most of the laser energy deposited in the cluster drives the ion explosion.

This result raises the interesting possibility that moderate-density plasmas with very high ion temperatures could be created from the intense illumination of a gas of clusters. Because each cluster in the medium explodes isotropically, the plasma that results after all the clusters in the gas have exploded will contain hot ions with an isotropic velocity distribution. Such novel, high-ion-temperature plasmas may find application in a host of experiments²⁷. For example, a plasma formed from the explosion of deuterium and tritium clusters may make possible table-top fusion experiments²⁸. Even a small laser (such as the one used in our experiments which delivers $\sim 100 \text{ mJ}$) could be used to produce a plasma, in a clustering DT gas jet, with an ion temperature of $\sim 10 \text{ keV}$ and a density of $\sim 10^{19} \text{ cm}^{-3}$ over an appreciable focal volume ($\sim 10^{-5} \text{ cm}^3$). Given that the expected disassembly time of this plasma would be⁵ of the order of 100 ps, such a plasma would exhibit an easily observable fusion yield of roughly 10^6 – 10^7 neutrons per shot²⁹. □

Received 26 September 1996; accepted 20 January 1997.

1. Castleman, A. W. & Keese, R. G. Gas-phase clusters: spanning the states of matter. *Science* **241**, 36–42 (1988).
2. McPherson, A., Thompson, B. D., Borisov, A. B., Boyer, K. & Rhodes, C. K. Multiphoton-induced X-ray emission at 4–5 keV from Xe atoms with multiple core vacancies. *Nature* **370**, 631–634 (1994).
3. Ditmire, T., Donnelly, T., Falcone, R. W. & Perry, M. D. Strong X-ray emission from high temperature plasmas produced by intense irradiation of clusters. *Phys. Rev. Lett.* **75**, 3122–3125 (1995).
4. Shao, Y. L. *et al.* Multi-keV electron generation in the interaction of intense laser pulses with Xe clusters. *Phys. Rev. Lett.* **77**, 3343–3346 (1996).
5. Ditmire, T., Donnelly, T., Rubenchik, A. M., Falcone, R. W. & Perry, M. D. The interaction of intense laser pulses with atomic clusters. *Phys. Rev. A* **53**, 3379–3402 (1996).
6. Cornaggia, C., Schmidt, M. & Normand, D. Coulomb explosion of CO₂ in an intense femtosecond laser field. *J. Phys. B*, **27**, L123–L130 (1994).
7. Perry, M. D. & Mourou, G. Terawatt to petawatt subpicosecond lasers. *Science* **264**, 917–924 (1994).
8. Augst, S., Meyerhofer, D. D., Strickland, D. & Chin, S. L. Laser ionization of noble gases by coulomb-barrier suppression. *J. Opt. Soc. B* **8**, 858–867 (1991).
9. Codling, K. & Frasinski, L. J. Coulomb explosion of simple molecules in intense laser fields. *Contemp. Phys.* **35**, 243–255 (1994).
10. Murnane, M. M., Kapteyn, H. C., Rosen, M. D. & Falcone, R. W. Ultrafast X-ray pulses from laser-produced plasmas. *Science* **251**, 531–536 (1991).
11. McPherson, A. *et al.* Multiphoton induced X-ray emission from Kr clusters on M-shell and L-shell transitions. *Phys. Rev. Lett.* **72**, 1810–1813 (1994).
12. Purnell, J., Snyder, E. M., Wei, S. & Castleman, A. W. Ultrafast laser-induced coulomb explosion of clusters with high charge states. *Chem. Phys. Lett.* **229**, 333–339 (1994).
13. Gordon, S. P., Donnelly, T., Sullivan, A., Hamster, H. & Falcone, R. W. X-Rays from microstructured targets heated by femtosecond lasers. *Opt. Lett.* **19**, 484–486 (1994).
14. Shepherd, R. *et al.* Characterization of short pulse laser-produced plasmas. *J. Quant. Spectrosc. Radiat. Transfer* **51**, 357–360 (1994).
15. Brégnacac, J. & Connerade, J. P. The giant resonance in atoms and clusters. *J. Phys. B*, **27**, 3795–3828 (1994).
16. Nagata, T., Hirokawa, J. & Kondo, T. Photodissociation of Ar₂⁺ cluster ions. *Chem. Phys. Lett.* **176**, 526–528 (1991).
17. Fraser, D. J. & Hutchinson, M. H. R. High intensity titanium-doped sapphire laser. *J. Mod. Opt.* **43**, 1055–1062 (1996).
18. Strickland, D. T., Beaudoin, Y., Dietrich, P. & Corkum, P. B. Optical studies of interially confined molecular iodine ions. *Phys. Rev. Lett.* **68**, 2755–2758 (1992).
19. Snyder, E. M., Buzza, S. A. & Castleman, A. W. Intense field-matter interactions: multiple ionization of clusters. *Phys. Rev. Lett.* **77**, 3347–3350 (1996).
20. Gitomer, S. J. *et al.* Fast ions and hot electrons in the laser-plasma interaction. *Phys. Fluids* **29**, 2679–2688 (1986).
21. Decoste, R. & Ripin, B. H. High-energy expansion in laser plasma interactions. *Phys. Rev. Lett.* **40**, 34–37 (1978).
22. Wickens, L. M. & Allen, J. E. Free expansion of a plasma with two electron temperatures. *J. Plasma Phys.* **22**, 167–185 (1979).
23. Meyerhofer, D. D. *et al.* Resonance absorption in high-intensity contrast, picosecond laser-plasma interactions. *Phys. Fluids B*, **5**, 2584–2588 (1993).
24. Perry, M. D., Darrow, C., Coverdale, C. & Crane, J. K. Measurement of the local electron density by means of stimulated Raman scattering in a laser produced gas-jet plasma. *Opt. Lett.* **17**, 523–535 (1992).
25. Ditmire, T., Smith, R. A., Tisch, J. W. G. & Hutchinson, M. H. R. Absorption of intense laser pulses by gases of atomic clusters. *Phys. Rev. Lett.* (submitted).
26. Ditmire, T., Gumbrell, E. T., Smith, R. A., Mountford, L. & Hutchinson, M. H. R. Supersonic ionization wave driven by radiation transport in a short-pulse laser-produced plasma. *Phys. Rev. Lett.* **77**, 498–501 (1996).
27. Rose, S. J. High-power laser-produced plasmas and astrophysics. *Laser Part. Beams* **9**, 869–880 (1991).
28. Lindl, J., McCrory, R. L. & Campbell, E. M. Progress toward ignition and burn propagation in inertial confinement fusion. *Phys. Today* **45**, 32–40 (1992).
29. Teller, E. *Fusion* 4–12 (Academic, New York, 1981).

Acknowledgements. We thank P. Ruthven for technical assistance and P. L. Knight for helpful conversations and advice.

Correspondence should be addressed to T.D. (e-mail: t.ditmire@ic.ac.uk).

Assembly of ordered colloidal aggregates by electric-field-induced fluid flow

Syun-Ru Yeh^{*†}, Michael Seul^{††} & Boris I. Shraiman[§]

^{*} Department of Physiology & Biophysics, Albert Einstein College of Medicine, The Bronx, New York 10461, USA

^{††} BioArray Solutions, Waksman Institute, Old Hoes Lane, Piscataway, New Jersey 08855-0759, USA

[§] Bell Laboratories, Lucent Technologies, 700 Mountain Avenue, Murray Hill, New Jersey 07974, USA

Suspensions of colloidal particles form a variety of ordered planar structures at an interface in response to an a.c. or d.c. electric field applied normal to the interface^{1–3}. This field-induced pattern formation can be useful, for example, in the processing of materials. Here we explore the origin of the ordering phenomenon. We present evidence suggesting that the long-ranged attraction between particles which causes aggregation is mediated by electric-field-induced fluid flow. We have imaged an axially symmetric flow field around individual particles on a uniform electrode surface. The flow is induced by distortions in the applied electric field owing to inhomogeneities in the 'double layer' of ions and counterions at the electrode surface. The beads themselves can create these inhomogeneities, or alternatively, we can modify the electrode surfaces by lithographic patterning so as to introduce specified patterns into the aggregated structures.

Our experiments were performed in planar, rectangular wells, fabricated by sandwiching ~50- μm polymer (Kapton) spacers between a bottom electrode of oxide-capped silicon, SiO_x , and a top electrode of indium tin oxide, ITO (SiO_x/ITO); a pair of indium tin oxide electrodes (ITO/ITO) was also used in some cases. In this simple two-electrode arrangement, currents were monitored (in constant-voltage mode) by a potentiostat. Carboxylated (anionic) and aminated (cationic) polystyrene beads of 1 μm or 2 μm diameter, and with a titrated (total) surface charge density of ~50 $\mu\text{C}/\text{cm}^2$, were obtained from Molecular Probes (Eugene, OR). Polystyrene beads in the range 2–20 μm were obtained from Polysciences (Warrington, PA). Beads were suspended in aqueous sodium azide solution and observed in epifluorescence or reflection differential interference contrast.

The photographs in Fig. 1 depict planar aggregates of beads adjacent to a uniform SiO_x electrode in configurations of increasing internal order: gaseous (Fig. 1a), liquid (Fig. 1b) and crystalline (Fig. 1d) states, as well as a possible hexatic state (Fig. 1c), resembling that identified in other planar systems⁴, are readily identified. These stationary, dissipative states are separated by reversible transitions³. In ITO/ITO cells, transitions occur in response to a varying d.c. voltage whereas, in SiO_x/ITO (as well as in ITO/ITO cells), they occur in response to a varying a.c. voltage (at typical frequencies of ≤ 1 kHz). Under conditions pertinent to Fig. 1, the crystalline state appears at a typical voltage gradient of 0.025 $\text{V}/\mu\text{m}$. Significantly, in cells formed with an insulating thermal oxide (SiO_x) electrode, an applied d.c. voltage produces no effect. Cationic and anionic beads behave in essentially identical fashion. When the applied voltage is removed, the patterns disappear and, unless constrained by gravity, beads redistribute into the bulk.

Remarkably, aggregates on uniform surfaces spontaneously form large-scale structures in two distinct morphologies. Under the same

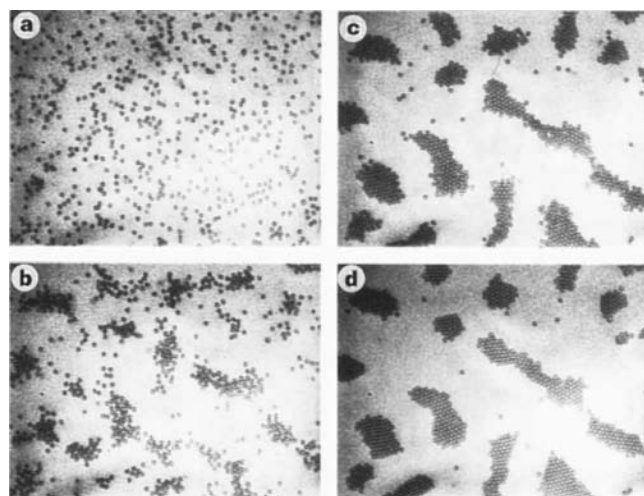


Figure 1 Sequence of increasingly ordered states of colloidal aggregates. **a**, gas (peak-to-peak voltage, $V_{pp} = 0$ V); **b**, liquid ($V_{pp} = 1.0$ V); **c**, hexatic(?) ($V_{pp} = 1.4$ V); **d**, crystalline ($V_{pp} = 1.5$ V). All transitions are reversible. Observations were made in an SiO_x/ITO cell (uniform electrodes) with a ~60- μm gap in response to an a.c. electric field of 1 kHz: the diameter of the carboxylated polystyrene beads is 2 μm . Images were obtained by reflection differential interference microscopy.

conditions as those producing the crystalline state (Fig. 1d), a network of finite elongated patches or interconnected bands is simultaneously observed on large scales (Fig. 2a)¹. In response to increasingly higher fields (≥ 0.04 $\text{V}/\mu\text{m}$, at typically several kilohertz), the crystalline state gives way to a configuration composed of mutually repelling particle clusters. Both morphologies display a characteristic scale, Λ , which exceeds the particle diameter, D , by an order of magnitude and does not vary significantly when the electrode spacing ('gap'), L , is varied in the range $5D \leq L \leq 40D$. For the pattern of bands (Fig. 2a), Λ corresponds to the 'mesh size' whereas for the pattern of clusters (Fig. 2b), it corresponds to the lattice parameter of the ordered 'superlattice' formed by the cluster centroids. Small clusters assume a characteristically faceted shape which depends on the number, n , of beads in the cluster (Fig. 2c). Furthermore, they reveal an intriguing property: the nearest-neighbour distance of beads in a cluster decays rapidly with n , attaining its 'bulk' value for $n \geq 7$.

Similar patterns in the form of meandering stripes and circular domains have been observed in monomolecular films of certain amphiphiles⁵. The formation of macroscopic patterns in such Langmuir films is known to be governed by competing attractive and dipolar repulsive interactions whose balance sets the scale Λ (ref. 6). A standard lattice model of such systems with competing interactions reproduces the faceted shapes favoured by small clusters⁷.

As with Langmuir films, the repulsive interaction between beads in our colloidal arrays is likely to be dipolar. Specifically, dipoles appear as a result of the polarization of the bead double layer in the external field^{8,9}. Direct evidence of a dipolar interaction is furnished by our observation of vertical strings of beads in high (~0.4 $\text{V}/\mu\text{m}$, 1 MHz) fields. The same dipolar polarization would also account for the attraction of the beads to the electrode double layer, a region characterized by a high electric field. Importantly, this dielectrophoretic force would operate in the a.c. as well as in the d.c. regime.

We attribute the attractive interaction leading to clustering of beads in the plane of the electrode to electroosmotic flow, which can arise as a result of spatial perturbations in the lateral current

[†] Previous address: AT&T Bell Laboratories, Murray Hill, New Jersey 07974, USA.

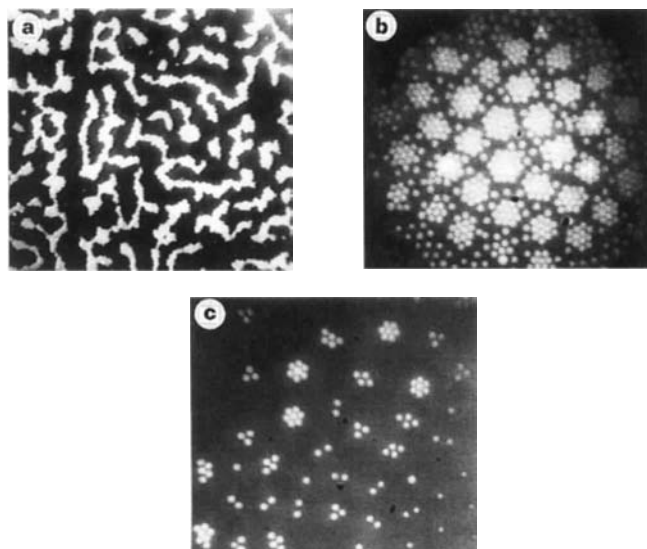


Figure 2 Large-scale morphologies displayed by colloidal aggregates on a uniform electrode. **a**, Network pattern of bands of (unresolved) beads observed under conditions favouring particle aggregation (see Fig. 1); the typical width of the bands corresponds to ~ 10 particle diameters. **b**, High-density (fluctuating) clusters, in a regime of large bead polarization ($V_{pp} \approx 0.2 \text{ V } \mu\text{m}^{-1}$). **c**, Faceted clusters observed at low coverage in high a.c. fields ($V_{pp} \approx 0.1 \text{ V } \mu\text{m}^{-1}$). The bead diameter is $2 \mu\text{m}$; images were obtained by epifluorescence microscopy.

distribution introduced by the colloidal beads themselves or by explicit electrode patterning.

To investigate the possible role of fluid flow in the formation of colloidal arrays¹, small beads ($1 \mu\text{m}$) were employed as tracers to image the flow field associated with individual, large ($20 \mu\text{m}$) beads resting on the bottom electrode. The flow field is toroidal, with closed stream lines and with a radial component focused on the large bead (Fig. 3); typical velocities are estimated to be in the range of $1\text{--}10 \mu\text{m s}^{-1}$. Irrespective of the sign of the applied a.c. voltage, in-flow occurs predominantly along the electrode surface. Recirculation produces vertical fluid motion, but the large bead remains close to the electrode.

The flow acts on an individual bead (of radius R) with a drag force, $F_s \approx 6\pi\eta Rv$, η and v respectively denoting viscosity and fluid velocity. In the vertical direction, this force must be balanced by dielectrophoresis and (for large beads) gravity moving beads towards the electrode. In the plane of the electrode, the interaction of the flows associated with two adjacent beads can produce an attractive interaction. That is, a stationary 'test' bead positioned in the flow of another bead at a distance r would experience an attractive force due to drag. The flow field of each bead is approximately dipolar (Fig. 3), and we consequently expect the force to decay as r^{-3} . Although we have not visualized the flow in a cluster of beads directly, our estimate suggests that the relevant Reynolds number is small, $\nu\rho D/\eta \ll 1$, so that the velocity fields of individual beads will add linearly. As illustrated for a set of three beads in Fig. 3, the vectorial addition of the corresponding forces yields a resulting force directed toward the centroid of the set. Accordingly, we suggest that the drag forces generated by converging lateral flows close to the electrode are essential in producing the effective attraction which counterbalances the dipolar repulsion and leads to the formation of bead clusters.

To address the question as to how an applied d.c. or a.c. electric field generates the observed flow, we investigated the connection between field-induced flow and the local electrochemical properties

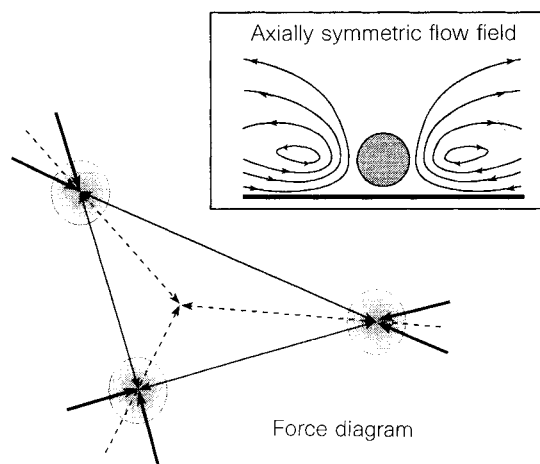


Figure 3 Illustration of the mutual advection of beads in their respective flow fields. Pairs of forces acting along lines connecting pairs of particles are shown, along with the resultant force toward the centroid. Inset, sketch (in side view) of toroidal flow field, as observed in the vicinity of an individual bead.



Figure 4 Effect of lithographic patterning of SiO_x electrode. Crystalline arrays of beads, formed in response to an applied a.c. voltage, are confined to etched regions of the surface, characterized by thin oxide. Note that the inter-stripe spacing of the pattern is not constant across the field of view. The bead diameter is $1 \mu\text{m}$. Images were obtained by epifluorescence microscopy.

of spatially patterned SiO_x electrodes. In the example of Fig. 4, photolithography served to introduce a stripe pattern by etching a thermal oxide layer. We find that, on such topographically patterned electrodes, beads assemble in regions of lower oxide thickness and correspondingly higher local current density; initial observations (not shown) indicate that suitable illumination of the semiconductor electrode can produce similar effects. This observation is consistent with additional experiments on patterned SiO_x in which we employed small tracer particles to facilitate the direct visualization of fluid flow to regions of high current density. These results point to variations in the spatial current distribution as a critical factor in producing flow.

We account for our observations by invoking a model based on electro-osmotic flow. As with electrokinetic phenomena in general, electro-osmotic flow originates in the action of an electric field on the double layer associated with a charged electrode surface

Predominance of vertical loss of carbon from surface waters of the equatorial Pacific Ocean

Dennis A. Hansell, Nicholas R. Bates & Craig A. Carlson

Bermuda Biological Station for Research, Inc., St Georges, GE-01, Bermuda

The equatorial Pacific Ocean makes a significant contribution to global carbon fluxes through both degassing of CO₂ to the atmosphere and new primary production^{1–4}; the eastern and central region is the source of most of the 1–2 Pg (10¹⁵ g) of CO₂ supplied annually to the atmosphere by the equatorial oceans⁵, and new primary production in the region may account for up to 18–56% of this global oceanic value⁶. The fate of carbon fixed by new primary production—whether removed to the deep ocean as sinking particles or retained in surface waters requires critical assessment because of the very different timescales of C removal that each process entails. Here we evaluate the transformations of carbon and nitrogen compounds in the surface waters of the South Equatorial Current of the Pacific Ocean. We calculate that carbon removed from the surface layer by degassing and sinking organic particles accounted for 41% and 53%, respectively, of the total C depletion during boreal autumn, 1992. The net accumulation of organic matter in the surface layer, a precondition for its eventual transport away from the Equator by horizontal advection, accounted for <6% of the drawdown, in substantial disagreement with the values up to 75% estimated from recent studies^{7–9}.

Data for this analysis were taken from published reports^{9–11}. Concentrations of total carbon dioxide (referred to as dissolved inorganic carbon, DIC), total organic carbon (TOC), nitrate plus nitrite (referred to as total inorganic nitrogen, TIN), and total organic nitrogen (TON) were determined as part of the US National Oceanographic and Atmospheric Administration OACES program during boreal autumn, 1992, and during boreal spring, 1994. Concentrations of total organic carbon and nitrogen include both the suspended particulate and dissolved fractions, and generally exclude rapidly sinking particles that contribute to export. The data evaluated here were collected in the surface layer (<40 m depth) of the South Equatorial Current (SEC) along three meridians (110° W, 125° W, and 140° W) during non-El Niño conditions. Earlier evaluations of the SEC during this period indicate that the uniform westward flow of this system lends itself well to mass-balance calculations such as performed here¹².

Profiles of carbon and nitrogen on the equator show decreases in the concentrations of the inorganic species with decreasing depth, while organic species increase in concentration (Fig. 1). These trends indicate the exchange of material between the pools due to biological activity as water upwells into the euphotic zone. The source water for upwelling, the Equatorial Undercurrent, lies at depths of 50–100 m in these figures¹. DIC decreased in concentration by ~70 µm kg⁻¹ from a depth of 50 m to the surface (Fig. 1a), with carbon loss attributable to degassing of CO₂ to the atmosphere, downward flux of sinking particles, accumulation of suspended TOC, and, potentially, dilution with low-DIC water. Nitrate decreased by 12 µM in the surface 50 m (Fig. 1b), with loss attributable to the same processes that account for carbon removal, except for degassing to the atmosphere.

To estimate the accumulation of mass into the TOC and TON pools, property–property plots were employed. This analytical approach overcomes the potential effects of dilution on the inorganic concentrations. There was a good correlation between TIN and TON concentrations (Fig. 2a). The slope of the regression

immersed in an electrolyte^{8,9}. In contrast to the standard geometry of planar electro-osmosis¹⁰, the applied field in our experiments is directed normal to the electrode surface. Nevertheless, any spatial modulation in the interfacial electrochemical potential suffices to generate lateral fluid flow. Specifically, a feature of transverse dimension R introduces a transverse variation, $\nabla_{\perp} J \approx J_0/R$ in the (ionic) current distribution, J_0 . The corresponding potential gradient acts on mobile charges in the Debye layer of the electrode, and, as shown in our experiments with patterned surfaces, the resulting electro-osmotic flow is directed toward regions of high current density.

We derive an order-of-magnitude estimate of the fluid velocity as follows. For simplicity, we assume that within the electrode double-layer ionic current is carried by a single species of (excess) charge e and concentration n . The viscous balance governing the solvent flow within the Debye layer of the electrode is: $\eta \nabla \mathbf{v} - \nabla p - en \nabla \Phi = 0$, $\mathbf{v}$ representing the fluid velocity, Φ the electric potential, and p the pressure. The electrochemical potential, $\mu = kT \log n + e\Phi$, determines the current, $\mathbf{J} = -(enD/kT) \nabla \mu$, so that we may express $en \nabla \Phi$ in the form $en \nabla \Phi = -(kT/eD) \mathbf{J} - kT \nabla n$. This substitution yields the balance equation $\eta \nabla^2 \mathbf{v} - \nabla p + (kT/eD) \mathbf{J} = 0$. The pressure with osmotic correction, $\bar{p} = p - kTn$, is eliminated by taking a curl so that: $\eta \nabla^2 \nabla \times \mathbf{v} + (kT/eD) \nabla \times \mathbf{J} = 0$. The relevant vertical scale is set by the Debye length, l_D , and we approximate $\nabla^2 \nabla \rightarrow \partial_z^2 \partial_z \rightarrow l_D^{-3}$ and $\nabla \times \mathbf{J} \rightarrow \nabla_{\perp} J_z / e$, yielding for the transverse fluid velocity the expression $v_{\perp} \approx (k_B T l_D^2 / \eta D) \nabla_{\perp} J_z / e$. With $kT/\eta D \approx 4$ nm, and a Debye length $l_D \approx 300$ Å, a value of 10 mA cm⁻² for the current density gives a velocity of ~7.5 µm s⁻¹. Inspection of the balance equation shows that the fluid must flow in the direction of higher current density. Magnitude and direction of the fluid velocity are thus consistent with our flow decoration experiments.

We note several points. First, the velocity, v , is determined by variations in the number current, J/e . Assuming a 1 : 1 electrolyte, it is therefore independent of the sign of the predominant charge carrier and independent of the direction of the applied field. Provided that experimental frequencies are low compared to the characteristic frequency, $D/l_D^2 \approx 10$ kHz, the double layer recharges essentially instantaneously when the field is reversed. Second, our analysis suggests that the maximum velocity is attained in the limit $R \rightarrow l_D$ so that $|v|_{\max} \approx (kT/\eta D) l_D^2 (J_0/e)$. Third, because the field-induced force acting on the fluid is confined to the electrode double layer, we expect the fluid velocity to attain its maximal value in the vicinity of the interface. That is what we observe.

Our model addresses the situation where the current is modulated directly, that is by patterning of the electrode surface. The presence of a bead near the electrode surface has an analogous effect: the bead obstructs the flow of current and introduces a dipolar enhancement in the current density surrounding the bead. In accordance with our argument, we expect lateral flow in the boundary layer converging towards the bead, as observed. The flow pattern associated with dense-packed bead arrays remains to be examined more carefully. □

Received 4 October 1996; accepted 7 January 1997.

1. Richetti, F., Prost, J. & Barois, P. *J. Phys. Lett.* **45**, L1137–L1143 (1984).
2. Giersig, M. & Mulvaney, P. *Langmuir* **9**, 3408–3413 (1993); *J. Phys. Chem.* **97**, 6334–6336 (1993).
3. Trau, M., Saville, D. A. & Aksay, I. A. *Science* **272**, 706–709 (1996).
4. Seul, M. & Murray, C. A. *Science* **262**, 558–560 (1993).
5. Seul, M. & Chen, V. S. *Phys. Rev. Lett.* **70**, 1658–1661 (1993).
6. Seul, M. & Andelman, D. *Science* **267**, 476–484 (1995).
7. Hurley, M. M. & Singer, S. J. *J. Phys. Chem.* **96**, 1951–1956 (1992).
8. Dukhin, S. S. & Derjaguin, B. V. *Surface and Colloid Science* Vol. 7 (ed. Matijević, E.) (Wiley, New York, 1974).
9. Russell, W. B., Saville, D. A. & Schowalter, W. R. *Colloidal Dispersions* (Cambridge Univ. Press, 1989).
10. Anderson, J. L. *Annu. Rev. Fluid Mech.* **21**, 61–99 (1989).

Acknowledgements. We thank R. Wong for advice on silicon processing. M.S. acknowledges helpful conversations with A. Ajdari and M. Trau.

Correspondence and requests for materials should be addressed to M.S. (e-mail: moire@worldnet.att.net).

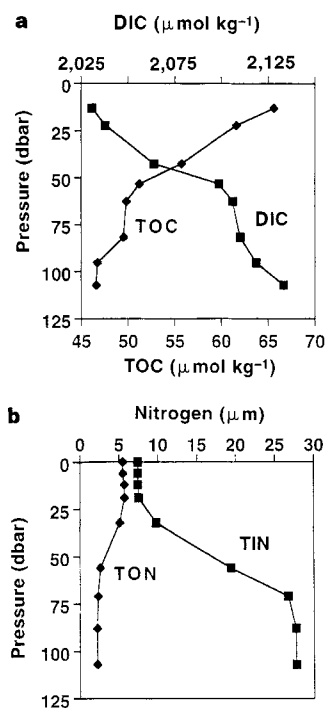


Figure 1 Profiles of **a**, dissolved inorganic carbon¹¹, DIC, and total organic carbon (TOC)¹¹, and **b**, total inorganic nitrogen (TIN)¹² and total organic nitrogen (TON)¹² in the central equatorial Pacific Ocean. Sampling locations and times were 0°30.8' N, 125° W (boreal autumn, 1992) and 1°18' S, 110° W (boreal spring, 1994) for carbon and nitrogen, respectively. The profiles shown are from different locations and times because concurrently determined data at high vertical resolution were not available. One decibar (dbar) is about the pressure due to 1 m depth of sea water.

indicates that an average of 27.1% (standard error of the slope s.e. = 0.010) of the inorganic nitrogen (TIN) removed from solution was found to accumulate as organic matter. The remainder of the nitrogen deficit (72.9%) must have been removed from the surface layer as sinking organic particles. There is no other loss term, such as air-sea exchange, to explain the removal of the inorganic fraction of nitrogen.

A similar property-property plot for DIC and TOC indicates that 4.4% (s.e. of slope, 0.018) of the total depletion in inorganic carbon can be explained by the accumulation of suspended TOC in the surface layer (Fig. 2b). Clearly there was a very wide range in the concentrations of DIC (~100 μmol kg⁻¹) and a relatively small range in TOC (~10 μmol kg⁻¹). As much as 20–30% of the range in TOC concentrations reported here may reflect the precision of the TOC measurement, hence a low correlation coefficient of -0.41.

The assessment of DIC drawdown that accumulates in the TOC pool as represented in Fig. 2b is complicated by two processes that have opposing impacts on the correlation shown. These processes are local upwelling of water with DIC:TOC ratios different than found at the surface, and DIC loss by the formation of CaCO₃ (calcification). The effect of local upwelling on the regression in Fig. 2b would be to increase the slope, as exemplified by a curve that would result from mixing surface water found along 110° W (open circles in Fig. 2b) with upwelling source water (50–65 m)¹ on the Equator (the triangle in Fig. 2b). The slopes from individual meridians decrease to the west, perhaps reflecting a lessening of local upwelling from the Equatorial Undercurrent as this current deepens to the west¹².

Given the effect of local upwelling alone, the fraction of DIC drawdown found in the TOC pool must be reported as being ≤4.4%

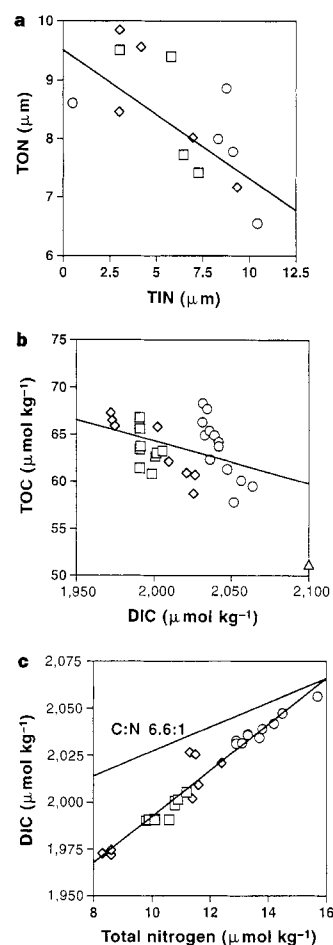


Figure 2 Correlations between **a**, mean concentrations of TIN and TON in the upper 40 m between the Equator and 10° S during boreal fall, 1992⁹ ($y = -0.271x + 10.0$; $r = -0.67$), and **b**, concentrations of DIC and TOC in the upper 25 m between the Equator and 4° S during boreal autumn, 1992¹¹ (data were normalized to a salinity of 34.8; $y = -0.044x + 152.0$; $r = -0.41$). The slopes from individual meridians in **b** were 0.268 (100° W), 0.125 (125° W), and 0.157 (140° W). The mean of the TOC and DIC concentrations at 50–65 m (two depths) on 110° W at the Equator, intended to indicate the concentrations in upwelling water at that site, are indicated (Δ). **c**, Concentrations of DIC¹¹ and estimated total nitrogen (estimated TON plus measured TIN) in the upper 25 m between the Equator and 4° S during boreal autumn, 1992 ($y = 12.39x + 1868$; $r = 0.97$). TON determinations from the surface layer are few, hence the need to estimate concentrations. The line predicted from the Redfield ratio¹⁴ loss of inorganic carbon (C:N 6.6:1) corresponding to the estimated loss of total nitrogen is shown. Data were taken along 110° W ($\circ$), 125° W ($\diamond$) and 140° W ($\square$). Data in **a** were taken from a wider zonal range because of the more limited data set available closer to the areas of equatorial upwelling. All data, except that reported in **a**, were normalized to a salinity of 34.8. r , Pearson product moment correlation coefficient; p , level of significance, <0.01 in all plots.

as the actual contribution of local upwelling is uncertain. Calcification, on the other hand, removes DIC without the production of organic matter so the effect is to decrease the slope in Fig. 2b. During this period, the proportion of CaCO₃ production in the SEC was 22% of soft-tissue formation¹³, so the slope could be underestimated by this fraction. Given the uncertainties of the net effect of calcification and local upwelling, the fraction of drawdown accounted for by TOC production must be <6%.

It is the fraction of the DIC depletion accumulating as TOC that can be exported from the site of upwelling by horizontal advection. The remainder of the inorganic carbon deficit must be accounted

for by degassing of CO_2 to the atmosphere and by downward flux of sinking particles. Estimates of the partitioning between these latter two processes can be made using the downward flux of organic nitrogen as a measure of the downward export of organic carbon. In order to estimate downward flux of nitrogen, and because concurrent measurements of TOC, DIC and TON near the Equator were few, we first estimated the amount of nitrogen that accumulated in the suspended organic fraction. TON concentrations were estimated from TOC concentrations measured during the OACES program by assuming a C:N atomic ratio of 12 for total organic matter, a value reported for the surface layer along 110°W (ref. 10). Measured DIC concentrations were then plotted against estimated concentrations of total nitrogen (TIN plus estimated TON; Fig. 2c; s.e. of slope, 0.57). A decrease in total nitrogen of $8\text{ }\mu\text{M}$ corresponded to a $99.4\text{ }\mu\text{mol kg}^{-1}$ decrease in DIC concentrations. The organic carbon removed to depth with the nitrogen was estimated using a C:N atomic ratio for sinking particles of 6.6. The downward flux of organic carbon represented $52.8\text{ }\mu\text{mol kg}^{-1}$ (53%) of the total DIC drawdown of $99.4\text{ }\mu\text{mol kg}^{-1}$. This analysis indicates that <6% and 53% of DIC depletion occurred owing to the production of TOC (with the potential for horizontal export by advection from upwelling system) and downward flux of sinking particles, respectively. Mass balance requires that the remaining 41% of inorganic carbon loss take place by degassing of CO_2 to the atmosphere.

TOC production and accumulation in the surface layer accounted for ~10% of the net community production (sum of TOC accumulation and downward flux of sinking particles) of carbon, compared to approximately 27% for nitrogen. If these data sets were directly comparable, we would expect the fractional accumulation of organic carbon in the surface layer to be greater than that for organic nitrogen. This follows from the high C:N ratio of the material. Such an inconsistency suggests that a direct comparison of the data sets is not warranted, perhaps due to analytical uncertainties and natural variability in the equatorial system.

Sensitivity analysis for the C:N ratios for sinking particles demonstrates that the proportion of C found in the TOC pool remains the same, while the contributions of export by degassing and sinking particles varies. By assuming a higher C:N ratio for sinking particles of 9, this contribution to export increases to 72%.

Assessing the fate of carbon fixed by new production is critical for determining the role of the biological pump in the long-term removal of carbon from the surface ocean. Sinking particles can remove organic carbon to below the main thermocline, thereby sequestering the carbon for periods approaching the timescale of oceanic turnover (10^2 – 10^4 years). Horizontal export of organic matter from the upwelling regions in the equatorial Pacific Ocean results in negligible removal of carbon from the surface ocean, leaving it available for export to the atmosphere on short timescales. The results presented here demonstrate that vertical processes (carbon transfer by degassing to the atmosphere and by downward flux of biogenic particles to the deep ocean) predominate in the export of carbon from this globally important region. □

Received 20 June 1996; accepted 13 January 1997.

1. Murray, J. W., Barber, R. T., Roman, M. R., Bacon, M. P. & Feely, R. A. *Science* **266**, 58–65 (1994).
2. Feely, R. A. *et al.* *J. Geophys. Res.* **92**, 6545–6558 (1987).
3. Francey, R. J. *et al.* *Nature* **373**, 326–330 (1995).
4. Wong, C. S., Chan, Y. H., Page, J. S., Smith, G. E. & Bellegay, R. D. *Tellus* **45**, 64–79 (1993).
5. Tans, P. P., Fung, I. Y. & Takahashi, T. *Science* **247**, 1431–1438 (1990).
6. Chavez, F. P. & Barber, R. T. *Deep-Sea Res.* **34**, 1229–1243 (1987).
7. Feely, R. A. *et al.* *Deep-Sea Res.* **42**, 365–386 (1995).
8. Peltzer, E. T. & Hayward, N. A. *Deep-Sea Res.* **43**, 1155–1180 (1996).
9. Libby, P. S. & Wheeler, P. A. *Deep-Sea Res.* **1** (in the press).
10. Hansell, D. A. & Waterhouse, T. Y. *Deep-Sea Res.* **1** (in the press).
11. Lamb, M. F. *et al.* *Data Rep. ERL PMEL-56* (NOAA, Seattle, WA, 1995).
12. Lukas, R. & Lundstrom, E. J. *Geophys. Res.* **96**, 3343–3357 (1991).
13. Wanninkhof, R. *et al.* *Deep-Sea Res.* **42**, 387–409 (1995).
14. Redfield, A. C., Ketchum, B. H. & Richards, F. A. in *The Sea* Vol. 2 (ed. Hill, M. N.) 26–77 (Wiley, New York, 1963).

Acknowledgements. We thank R. Feely, M. Roberts, E. Peltzer, S. Libby and P. Wheeler for their analytical work, and whose data we used in the analysis. Comments from R. Murnane were appreciated.

Correspondence should be addressed to D.A.H. (e-mail: dennis@bbsr.edu).

GPS measurements of present-day convergence across the Nepal Himalaya

Roger Bilham*, Kristine Larson†, Jeffrey Freymueller‡ & Project Idylhim members§

* CIRES and Department of Geological Sciences, University of Colorado, Boulder, Colorado 80309, USA

† Department of Aerospace Engineering Sciences, University of Colorado, Boulder, Colorado 80309, USA

‡ Geophysical Institute, University of Alaska, Fairbanks, Alaska 99775, USA

The high elevations of the Himalaya and Tibet result from the continuing collision between India and Asia, which started more than 60 million years ago^{1–4}. From geological and seismic studies of the slip rate of faults in Asia⁵, it is believed that approximately one-third of the present-day convergence rate between India and Asia ($58 \pm 4\text{ mm yr}^{-1}$) is responsible for the shortening, uplift and moderate seismicity of the Himalaya. Great earthquakes also occur infrequently in this region, releasing in minutes the elastic strain accumulated near the boundary zone over several centuries, and accounting for most of the advance of the Himalaya over the plains of India. The recurrence time for these great earthquakes is determined by the rate of slip of India beneath Tibet, which has hitherto been estimated indirectly from global plate motions⁶, from the slip rates of faults in Asia^{7,8}, from seismic productivity⁹, and from the advance of sediments on the northern Ganges plain¹⁰. Here we report geodetic measurements, using the Global Positioning System (GPS), of the rate of contraction across the Himalaya, which we find to be $17.52 \pm 2\text{ mm yr}^{-1}$. From the form of the deformation field, we estimate the rate of slip of India beneath Tibet to be $20.5 \pm 2\text{ mm yr}^{-1}$. Strain sufficient to drive one or more great Himalayan earthquakes, with slip similar to that accompanying the magnitude 8.1 Bihar/Nepal 1934 earthquake, may currently be available in western Nepal.

In March 1991 we used Global Positioning System (GPS) geodesy to measure the relative positions of 24 points in India, Nepal and Tibet. Some of these points were remeasured in 1992¹¹ and in 1995, and additional points were installed in November 1995 in collaborative programmes between US, Nepalese, Indian, Chinese and French teams. Data were obtained in five 8-h sessions in 1991 and 1992 and for three to five consecutive 24-h sessions in 1995. We occupied a central point near Kathmandu (NAGA) continuously during each survey relative to which all positions are referred. The data from all surveys were analysed with the GIPSY software¹², using an estimation strategy summarized in ref. 13, and orbital data from the global tracking network using a geodetic reference frame defined by ITRF94 (ref. 14).

For 20 points common to the 1994 and 1991 surveys the mean formal uncertainty in northward velocities relative to NAGA is $\pm 1.1\text{ mm yr}^{-1}$, and $\pm 1.8\text{ mm yr}^{-1}$ for eastward components, considerably smaller than previous GPS measurements in the Himalaya¹⁵. The maximum observed convergence velocity between northern India and southern Tibet is $18.2 \pm 2\text{ mm yr}^{-1}$ at $13^\circ \pm 4^\circ$. Mean velocity uncertainties relative to Bangalore, southern India (the nearest permanent site), are $\pm 1.9\text{ mm yr}^{-1}$ north and $\pm 2.3\text{ mm yr}^{-1}$ east (Fig. 1). In southern Nepal, convergence velocities

§ F. Jouanne (Univ. Savoie, Chambéry, France), P. Le Fort, P. Leturmy & J. L. Mugnier (IGCA-CNRS, UJF, Grenoble, France), J. F. Gamond, J. P. Glot & J. Martinod (LGIT-CNRS, UJF, Grenoble, France), N. L. Chaudury, G. R. Chitrakar, U. P. Gautam, B. P. Koirala, M. R. Pandey, R. Ranabhat, S. N. Sapkota, P. L. Shrestha, M. C. Thakuri, U. R. Timilsina & D. R. Tiwari (Dept Mines and Geology, Kathmandu, Nepal), G. Vidal (ENS Lyon, France), C. Vigny (ENS Paris, France), A. Galy (CRPG-CNRS, Nancy, France), B. de Voogd (Univ. Pau, France).

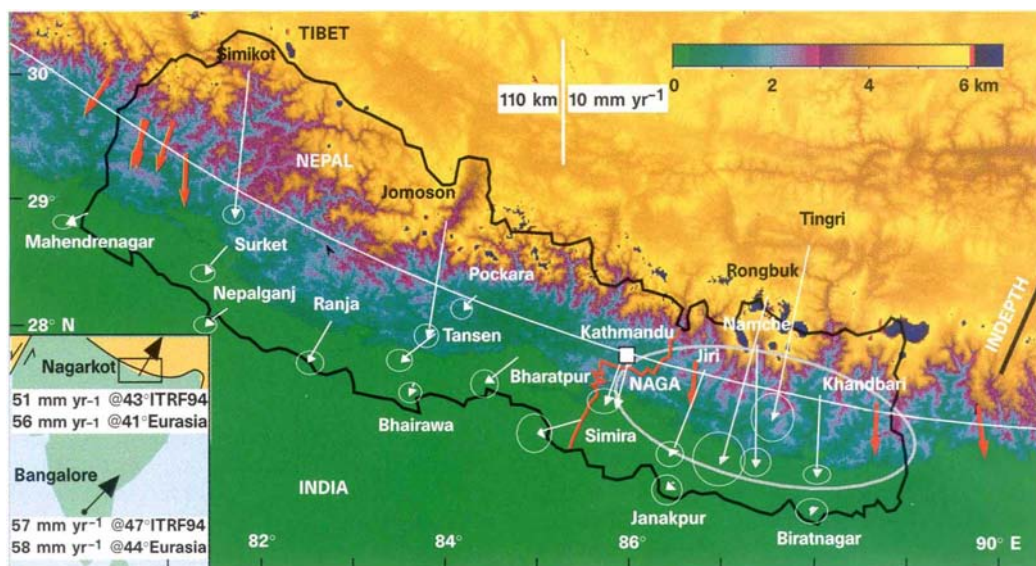


Figure 1 Himalayan GPS velocity vectors (white lines with arrowheads) 1991–95 relative to Bangalore. The levelling line (see text; red) passes near Nagarkot (white square, NAGA, to which GPS processing was referenced) and the western edge of the inferred rupture of the 1934 earthquake (large ellipse). White line, small-circle approximation for the Himalayan arc; black line, INDEPTH seismic profile²⁶;

red arrows, slip vectors for moderate earthquakes¹⁵ since 1960. Inset, location map showing Bangalore and Nagarkot vectors^{18,19} for ITRF94 and for a no-net-rotation frame for Eurasia fixed with uncertainties of 4 mm in scale and 5° in azimuth. Error ellipses are one standard deviation; the colour scale indicates elevation in kilometres.

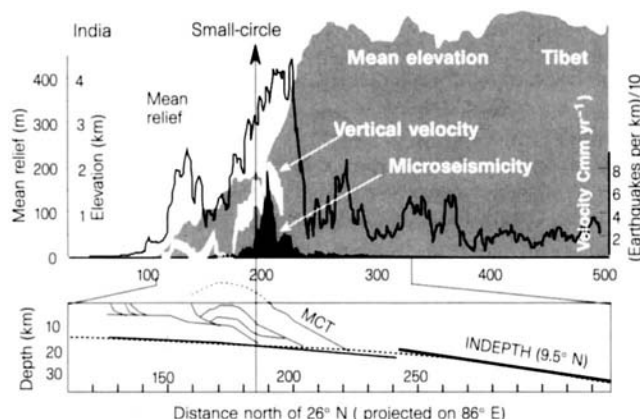


Figure 2 Top, levelling data (white envelope), seismic data²⁸ (black envelope), mean elevation data (shaded) and topographic relief (0.05° digital terrain data) at $86^\circ \pm 1^\circ$ E. The high peaks of the Himalaya are found close to the edge of the Tibetan plateau, 20–30 km north of the region of maximum relief. A small-circle fit to maximum Himalayan stream gradients and moderate seismicity²⁴ intersects the section near maximum downslope tilts ($2.5 \mu\text{rad yr}^{-1}$) in the vertical velocity field. Bottom, a typical geological section for eastern Nepal²⁵ (MCT, Main Central Thrust) and the estimated dip of the upper surface of the Indian plate from the INDEPTH seismic profile²⁶ (see Fig. 1), using the small-circle approximation for the Himalayan arc to project at 86° E. Dashed line, polynomial fit used to approximate flexure of the Indian plate used in Fig. 3.

are low but approximately arc-normal. In contrast, the mean convergence velocity for points on the southern margin of the Tibetan plateau is apparently directed at the approaching Indian plate ($17^\circ \pm 4^\circ$), arc-normal in eastern Nepal but trending 22° east of arc-normal in western Nepal, not significantly different from the slip vectors of published focal mechanisms for nearby thrust earthquakes¹⁶. Focal mechanisms for moderate earthquakes along the Himalayan arc approximate arc-normal slip which, if India is assumed rigid, requires $\sim 100^\circ$ E extension of the southern Tibetan plateau at a rate approximating the rate at which India collides with Tibet^{17–19}. Thus if extension occurs steadily, Simikot and Tingri should recede from each other at $\sim 7 \pm 2 \text{ mm yr}^{-1}$. That these points currently converge at $3 \pm 3 \text{ mm yr}^{-1}$ requires episodic and/or heterogeneous east–west extension of southern Tibet.

GPS data from NAGA and Bangalore were used recently to confirm²⁰ the NUVEL1 hypothesis that India rotates anticlockwise relative to the Australia plate⁶. The data we have since collected at Bangalore confirm NUVEL1 motions in southern India, but it is clear that NAGA, which approaches Bangalore at $5.0 \pm 1.5 \text{ mm yr}^{-1}$, is within the deforming zone between India and Eurasia, consistent with earlier findings that minor deformation occurs in the Lesser Himalaya²¹. The eastward motion of our

points in Nepal and south central Tibet relative to Eurasia, assuming no net rotation of India and Eurasia, is $25 \pm 5 \text{ mm yr}^{-1}$.

Because we may have incompletely sampled deformation in southern Tibet, our GPS data do not provide an upper bound to the rate at which India slips beneath Tibet. But, if we assume that the slip of India beneath Tibet can be approximated by slip on a subsurface dislocation (a mathematically convenient but by no means unique explanation for surface deformation), a strong constraint occurs because the maximum uplift rate in the vertical velocity field, and the maximum gradient in the horizontal velocity field, overlie the tip of an active subsurface dislocation^{22,23}. Moreover, if the dip of the dislocation is shallow ($<15^\circ$) the horizontal deformation field is approximately antisymmetric about the tip of this dislocation, permitting data from the south of the dislocation to yield approximate constraints on deformation to its north.

Vertical uncertainties in the GPS data are large ($>5 \text{ mm yr}^{-1}$), partly due to uncertainties in the phase centres of some of the antennas used, and these data are inadequate to constrain models of subsurface deformation. Recent spirit-levelling data (Fig. 2) obtained by the Nepalese Survey Department (1977–95), however, confirm a previously identified 40-km wavelength region of uplift in northern Nepal rising at a peak rate of $7 \pm 3 \text{ mm yr}^{-1}$ relative to the

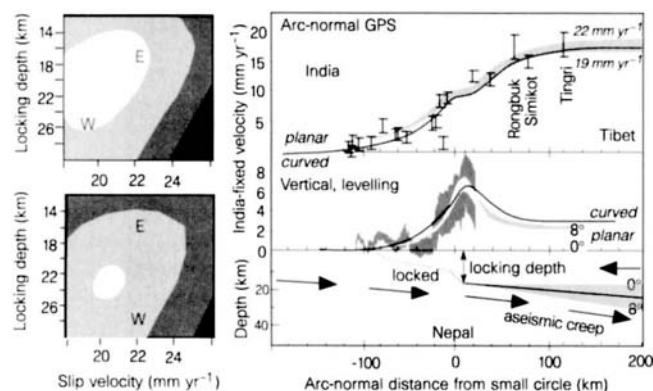


Figure 3 Right, surface deformation resulting from subsurface creep on planar and curved dislocations compared to GPS horizontal velocities and levelling data. Left, shaded regions correspond to 68%, 90% and 99% confidence intervals that satisfy synthetic and observed deformation for a planar dislocation dipping 4° northwards (upper left) and for a curved dislocation (lower left) steepening from 3° northwards at the locking depth to 9° northwards beneath southern Tibet. A deeper locking depth in western Nepal (W) than in eastern Nepal (E) is indicated by the GPS data (68% confidence open ellipses, left) if constraints from the vertical data are ignored.

Indian border. Interestingly, the location of peak uplift corresponds to a maximum in arc-normal Himalayan relief (Fig. 2). The central segment of the Himalayan arc follows a 1,696-km-radius small-circle defined by a belt of moderate Himalayan earthquakes and maximum river gradients²⁴ which intersects the levelling line ~ 15 km south of this peak uplift. The similarity in wavelength of the mean arc-normal relief and vertical deformation field, and the correspondence between maximum southward tilt ($2.5 \mu\text{rad yr}^{-1}$) and maximum stream gradients, suggests that the deformation field we observe is not a transient feature of Himalayan uplift as we might suspect from the brief 18-yr duration of these interseismic measurements. One of several mechanisms that could account for the observed coincidence is that the region may be repeatedly taken close to its elastic limit, with average uplift equal to half its pre-seismic peak value of 2–5 m (assuming $7 \pm 3 \text{ mm yr}^{-1}$ for, say, 300–500 yr). If a fraction of this uplift were non-recoverable inelastic deformation, permanent uplift of the region would mimic the form of the interseismic deformation field.

To reconcile the levelling data with the GPS data we projected both onto a cross-section parallel to the mean convergence direction at 86°E , using the above small-circle approximation to correct for the curvature of the Himalaya (Fig. 1). The projected horizontal and vertical deformation data were then compared to the surface deformation accompanying slip on a planar dislocation (the simplest of possible slip geometries) locked below, and south of, the region of maximum uplift. A dislocation dipping at $4^\circ \pm 4^\circ$ to the north best fits the data with a slip rate of $20.5 \pm 2 \text{ mm yr}^{-1}$ and a locking depth of $20 \pm 4 \text{ km}$ (Fig. 3). When the vertical deformation data are ignored, independent best fits to the GPS data from eastern Nepal favour slip at $22 \pm 3 \text{ mm yr}^{-1}$ starting at $17 \pm 4 \text{ km}$ depth dipping to the north at $4^\circ \pm 4^\circ$, but the data from western Nepal require deeper locking depths ($24 \pm 4 \text{ km}$). Whereas slip on a curved dislocation surface dipping increasingly to the north (approximating the flexure of a descending Indian plate) is permitted by the data, we can exclude the possibility of active creep on dislocations at 10–20 km depths dipping to the north at 10° – 30° , on which moderate seismicity has been observed in the past century. Satisfactory models for a curved dislocation have a slip velocity of $20.5 \pm 1 \text{ mm yr}^{-1}$ and starting dip of $4^\circ \pm 2^\circ$ to the north (compare ref. 25) at a locking depth of $21 \pm 2 \text{ km}$, and increase smoothly in dip to 9° northward at a distance of 150 km north of the locking depth (compare ref. 26).

More complex geometries and slip distributions, such as those as have been proposed to explain the vertical velocity data²¹ cannot be constrained uniquely by our sparse horizontal GPS data, although we can exclude significant creep on surfaces dipping more than 9° northward. Moreover, convergence south of the Himalayan boundary thrusts in western Nepal (Fig. 1) and its absence in eastern Nepal reveals that deformation is not identical in both regions, indicating spatial and/or temporal variations in the convergence process. Synforms and antiforms with wavelengths of less than 100 km imply along-arc heterogeneity^{25,27}, but microseismic data in the Himalaya that might assist our identification of active subsurface geometries are insufficiently detailed except for the region near Kathmandu²⁸. A deep focus magnitude $M = 6.5$ earthquake occurred in 1988 in southeastern Nepal^{29,30} near the southern rupture limit of the great 1934 Bihar/Nepal earthquake. Although this event was too small to produce regionally significant surface deformation, viscoelastic relaxation of the region following the 1934 earthquake may partly be responsible for this event, and for observed differences in deformation style in eastern and western Nepal. Numerical estimates of viscous processes are hindered by incomplete information on the 1934 rupture geometry and slip distribution³¹.

The locking depth, an arc-parallel line in our two-dimensional model for sub-Tibetan slip, may correspond to the northern limit of rupture during the great 1934 Bihar/Nepal event. Assuming that rupture does not propagate south of the main frontal thrusts of the southern Himalaya³², or significantly north of this locking depth, the down-dip width of the 1934 rupture zone is estimated to be less than 90 km. This is 20% shorter than previous estimates, requiring estimates of co-seismic slip and renewal time based on seismic energy release from these events to be increased accordingly. In western Nepal there have been no recent great earthquakes for 200 years and perhaps for much longer. Earthquakes in 1803 (Uttarkashi) and 1833 (Nepal) have been assigned maximum magnitudes of $M = 7.7 \pm 0.2$ based on available historical data^{9,33}. Magnitude estimates for earlier events are unavailable, although from readings of colonial and Indian archives it is considered unlikely that a $M > 8$ event beneath the Kumaun Himalaya or western Nepal would have escaped mention in the past 350 years. Hence, one or more great earthquakes in western Nepal, if they occurred today, would release $> 6 \text{ m}$ (refs 33, 34) of co-seismic displacement, similar to other $M > 8$ Himalayan events that have occurred in the past 100 years. □

Received 31 July 1996; accepted 23 January 1997.

1. Molnar, P. *Annu. Rev. Earth Planet. Sci.* **12**, 489–518 (1984).
2. Molnar, P. *J. Himalayan Geol.* **1**, 131–154 (1990).
3. Harrison, M. T., Copeland, P., Kidd, W. S. F. & Yin, A. *Science* **255**, 1663–1670 (1992).
4. Ni, J. F. *Proc. Ind. Acad. Sci. (Earth Planet. Sci.)* **98**, 71–89 (1989).
5. Molnar, P. & Tapponnier, P. *Science* **189**, 449–462 (1975).
6. DeMets, C., Gordon, R., Argus, D. & Stein, S. *Geophys. J. Int.* **101**, 425–478 (1990).
7. Molnar, P. & Tapponnier, P. *J. Geophys. Res.* **83**, 5361–5375 (1978).
8. Avouac, J.-P. & Tapponnier, P. *Geophys. Res. Lett.* **20**, 895–898 (1993).
9. Molnar, P. & Deng, Q. *J. Geophys. Res.* **89**, 6203–6228 (1984).
10. Lyon Caen, H. & Molnar, P. *Tectonics* **4**, 513–518 (1985).
11. Jackson, M. & Bilham, R. *Geophys. Res. Lett.* **21**, 1169–1172 (1994).
12. Lichten, S. M. & Border, J. S. *J. Geophys. Res.* **92**, 12751–12762 (1987).
13. Larson, K. M. & Freymueller, J. *Geophys. Res. Lett.* **22**, 37–40 (1995).
14. Boucher, C., Altamimi, Z., Feissel, M. & Sillard, P. *Results and Analysis of the ITRF94 (IERS Tech. Note 20 IERS Central Bureau, Observatoire de Paris, 1996)*.
15. Anzidei, M. *Terra Nova* **6**, 82–89 (1994).
16. Molnar, P. & Lyon Caen, H. *Geophys. J. Int.* **99**, 123–153 (1989).
17. Armijo, R., Tapponnier, P., Mercier, J. L. & Tonglin, H. *J. Geophys. Res.* **91**, 13803–13872 (1986).
18. England, P. C. & Houseman, G. A. *J. Geophys. Res.* **91**, 3664–3676 (1986).
19. Molnar, P. & Tapponnier, P. *Geology* **5**, 212 (1977).
20. Freymueller, J. *et al. Geophys. Res. Lett.* **23**, 3107–3110 (1996).
21. Jackson, M. & Bilham, R. *J. Geophys. Res.* **99**, 13897–13912 (1994).
22. Savage, J. C. *J. Geophys. Res.* **88**, 4984–4996 (1983).
23. Okada, Y. *Bull. Seism. Soc. Am.* **75**, 1135–1154 (1985).
24. Seeber, L. & Gornitz, V. *Tectonophysics* **92**, 335–367 (1983).
25. Schelling, D. *Tectonics* **11**, 925–943 (1992).
26. Makovsky, Y., Klempner, S. L., Liyan, H., Deyuan, L. & Project INDEPTH team *Tectonics* **15**, 997–1005 (1996).
27. Johnson, M. R. W. *Tectonophysics* **239**, 139–147 (1994).
28. Pandey, M. R. *et al. Geophys. Res. Lett.* **22**, 751–754 (1995).
29. Pandey, M. R. & Nicolas, M. (Rep. 2, Dept. Mines and Geology, HMG Nepal, Kathmandu, 1988).

30. Banerjee, S. N. & Chakravarti, P. (eds) *Bihar-Nepal Earthquake 20 August 1988* (Spec. Publ. 31, Geol. Surv. India, Calcutta, 1993).
31. Yu, Tingto thesis, Univ. Colorado (1995).
32. Bilham, R., Bodin, P. & Jackson, M. *J. Nepal. Geol. Soc.* 11, 73–88 (1995).
33. Khatri, K. N. *Tectonophysics* 138, 79–92 (1987).
34. Molnar, P. *Ann. Geophys.* 5, 663–670 (1987).

Acknowledgements. We thank the Survey Department and the Ministry of Mines and Geology, HMG, Nepal, for their participation in the 1991 and 1995 measurements and for their generous contributions of new data. M. Jackson, B. Washburn and P. Molnar made substantial contributions to the initial GPS network. W. Wang, P. Bodin and E. Sigmundsson gathered GPS data in Tibet, and V. Gaur, J. Paul and S. Jade contributed data from India. R. Bürgmann, F. Blume, D. Shelling, P. Molnar, J.-P. Avouac and T. Herring provided insights into modelling and interpretation. This work was supported by the US NSF; we thank Trimble Navigation Solarex Corp. and Campbell Scientific Inc. for donations of equipment.

Correspondence should be addressed to R.G.B. (e-mail: bilham@stripe.Colorado.edu).

Confined subsurface microbial communities in Cretaceous rock

Lee R. Krumholz*, James P. McKinley†, Glenn A. Ulrich* & Joseph M. Suflita*

* Department of Botany and Microbiology, University of Oklahoma,
770 Van Vleet Oval, Norman, Oklahoma 73019, USA

† Pacific Northwest National Laboratory, Richland, Washington 99352, USA

Deep subsurface microbial communities¹ are believed to be supported by organic matter that was either deposited with the formation sediments or which migrated from the surface along groundwater flowpaths. Investigation has therefore focused on the existence of microorganisms in recently deposited or highly permeable sediments^{2,3}. Fewer reports have focused on consolidated rocks^{4–7}. These findings have often been limited by inadequate tracer methodology or non-sterile sampling techniques. Here we present evidence for the presence of spatially discrete microbial communities in Cretaceous rocks and advance a mechanism for the long-term survival of these subterranean communities. Samples were collected using aseptic methods and sensitive tracers⁸. Our results indicate that the main energy source for these communities is organic material trapped within shales. Microbial activity in shales appears to be greatly reduced, presumably because of their restrictive pore size⁹. However, organic material or its fermentation products could diffuse into adjacent, more permeable sandstones, where microbial activity was much more abundant. This process resulted in the presence of microbial communities at sandstone–shale interfaces. These microorganisms presumably ferment organic matter and carry out sulphate reduction and acetogenesis.

Alternating members of the Cretaceous Mancos (shale) and Dakota (sandstone) formations were drilled and sampled adjacent to Cerro Negro, a volcanic neck in central New Mexico near the village of Seboyeta. Core samples were collected into lexan liners using aqueous and particulate tracers for detection of potential contaminants⁹. The sampled strata represented relatively minor transgressive and regressive episodes within an overall transgressive sequence¹⁰. Sandstone members represented near-shore (shore face or near-shore bar) depositional environments; shale members represented adjacent, deeper and less energetic, near-shore ('toe-of-slope') environments.

A regional analysis of groundwater composition and hydrology using water from existing wells was conducted in parallel with this work (refs 9, 11, and J.P.McK., unpublished results). Recharge occurred on an upland mesa 10 km to the west and 425 m higher than the sampled well, where the water table was at a depth of 69 m. Cerro Negro groundwaters were sampled by setting downhole packers across sandstone units (shales were non-transmissive and would not yield water) and removing water to the surface using gas-

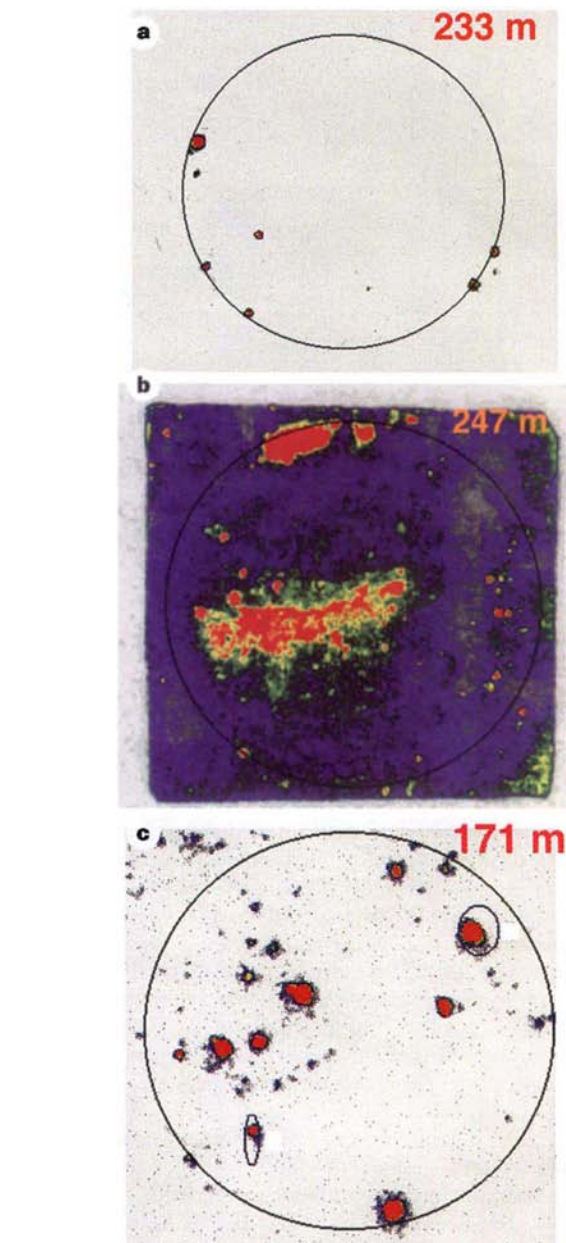


Figure 1 Radioisotope images of sulphate-reduction activity on the freshly exposed face of rock cores obtained from different depths at the Cerro Negro, New Mexico, exploratory drilling site. These images allow for localization of sulphate-reducing activity on a two-dimensional scale. **a**, 233-m core showing microbial contamination of the exterior surface. **b**, 247-m core showing generally high levels throughout, with maximal activity in the centre region. **c**, 171-m core showing patchy distribution of microbial activity. The interior (circled) represents the contact area of the cylindrical core with the foil. The level of radioactivity at that location on the foils is represented in colour (red > yellow > green > blue), although the scale is different for each foil. Small outlined areas in the 171-m core represent points of reference and not necessarily microbial activity.

driven Teflon bladder pumps. Dissolved oxygen was below detection (0.2 mg l^{-1}) for three sampled sandstone units, and sulphate and sulphide were present at $80\text{--}370 \text{ mg l}^{-1}$ and $8\text{--}14 \text{ mg l}^{-1}$, respectively. The sandstones and shales contained 0.1–2 wt% sulphide minerals (mainly pyrite). Analysis of isotope data for Paguate and Cubero sandstone aquifers indicated an age for these groundwaters of 20,000–30,000 years. Moreover, relatively low levels of Fe(III) were detected in the sediments (data not shown), and the

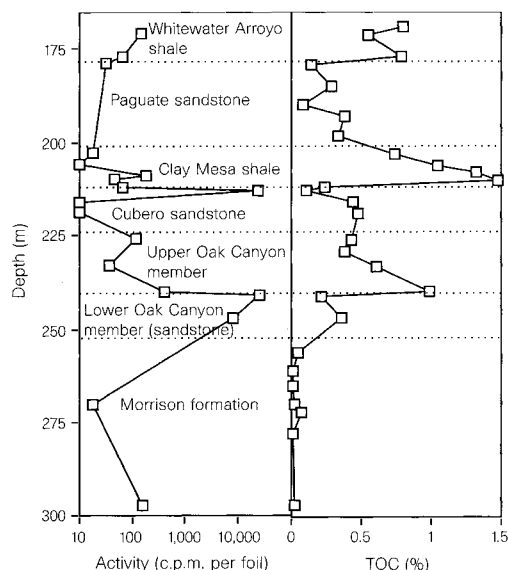


Figure 2 Plot of sulphate-reduction activity (left) and TOC (right) over a depth profile at the Cerro Negro site.

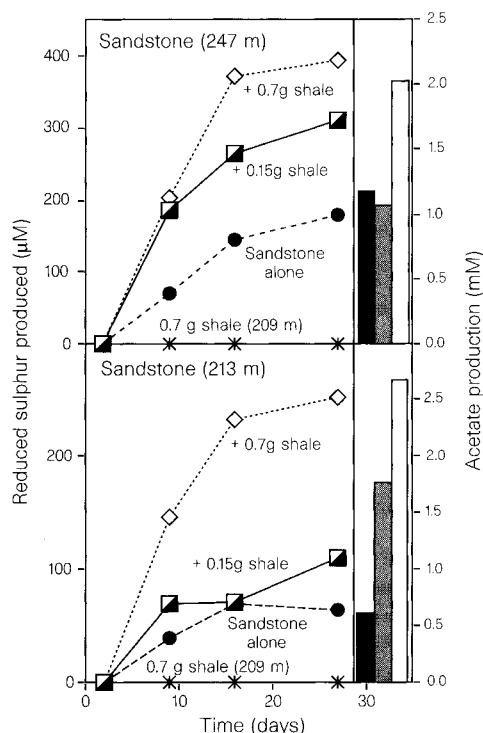


Figure 3 Plot of sulphate reduction as measured by the accumulation of Cr(II)-reducible sulphur. Black bars and circles represent sandstone incubations, hatched bars and half-filled squares represent sandstone and 0.15 g shale per tube, and the white bars and diamonds represent sandstone and 0.7 g shale per tube.

ground water contained $\sim 1 \text{ mg l}^{-1} \text{ NO}_3$ or less. Thus, we considered that bacterial sulphate reduction was a dominant respiratory process in these sediments.

Sulphate reduction was evaluated in freshly fractured cores using an *in situ* ^{35}S -sulphate reduction assay with a piece of oxidized silver foil as a ^{35}S -sulphide trap. This technique enabled us to visualize and quantify sulphate reduction in two dimensions, with only minimal disturbance of the indigenous rock microflora (Fig. 1). the foil

provided a fine-scale (1 mm resolution), two-dimensional image of sulphate reduction interior to the core. The distribution of ^{35}S -sulphide deposits relative to the core boundaries could be used as a self-tracing method for detecting drilling-induced microbial contamination. A few cores showed either no activity or greater activity around the periphery of the core (Fig. 1a), with the latter appearing to penetrate only 5 mm or less into the 6-cm core. Other cores showed relatively uniform and very high levels of sulphate reduction activity (Fig. 1b). In some cases, the activity appeared outside the core face, most probably due to diffusion of sulphide through the core walls to the foil. Other cores that showed much less activity tended to show a much more patchy distribution (Fig. 1c). These observations suggest that sulphate-reducing bacteria were limited to heterogeneously distributed microzones within these strata.

When plotted as a function of depth (Fig. 2a), several areas (at 213, 241 and 247 m depth) had 100 to 1,000 times more sulphate-reducing activity than the others. These regions of high sulphate-reducing activity, and presumably other microbial activities, were localized in sandstones and were generally at sandstone–shale interfaces. Within the Cubero sandstone, activity was highly localized at 213 m, with almost insignificant activity at locations only a few metres lower (216 and 219 m depth). In the case of the Lower Oak Canyon member, both of the samples available for analysis showed high activity, whereas the adjacent mudstone (shale) sampled from $\sim 1 \text{ m}$ above the highly active area (240 m) had very little activity. Stable-isotope analysis also indicated that sulphate reduction occurred in these rock formations. $\delta^{34}\text{S}$ values for Cr(II)-reducible sulphur extracted from CNAR samples 209, 213 and 247 were -14.9 , -26.6 and -16.3‰ , respectively, whereas the $\delta^{34}\text{S}$ value for Cubero sandstone groundwater sulphate was -9.0‰ . The preferential production of ^{34}S -depleted S^{2-} (ref. 12) is consistent with our hypothesis that biological sulphate reduction is an important geochemical process in this formation.

When the interfaces were examined in greater detail, significant gradients of total organic carbon (TOC) were measured (Fig. 2b). More TOC was present in the shales, with two areas of high sulphate-reduction activity in sandstone immediately beneath the most organic-rich strata at 210 and 237 m depth. These results indicate that shales may have served as reservoirs of electron donors for microorganisms living in sandstones near the interfaces.

We questioned why microbial activity should be mainly limited to sandstones and was not comparable in shales. Previous studies concluded that decreases in pore size restrict microbial penetration, reproduction and metabolic activity^{13,14}. It has been proposed that shale pore size is also a factor limiting microbial proliferation in this geological setting⁹, based on a comparison of microbial activity studies with pore size in the same rock samples. Median-pore throat diameter in the Clay Mesa Shale ranged from 10 to 30 nm, compared with 90 nm to 13 μm in the adjacent Cubero sandstone⁹.

To test whether shales supply electron donors for sulphate reduction in sandstones, we set up experiments in which ground rock material, taken from the innermost portion of cores, was incubated with filter-sterilized groundwater from the site. Figure 3 shows that addition of shale to a sandstone–groundwater slurry stimulated sulphate-reducing activity and that the increase was proportional to the amount of shale added (measured as Cr²⁺-reducible sulphur equivalents; 33–111 nmol sulphide (g rock)⁻¹ day⁻¹). With a reduced sulphur pool of 0.9 and 1.1 mg Sg⁻¹ rock (sandstones from 213 and 247 m depth), sulphide production at these measured rates would equal the sulphide pool in 1–4 years, so clearly these rates are overestimates of the actual *in situ* rate. No detectable sulphate reduction occurred when shales alone were incubated with groundwater. Sandstones from depths of 213 and 247 m reduced sulphate throughout the one-month incubation, but a similar experiment using sandstone from just 3 m lower (216 m depth) had no detectable activity. Addition of organic carbon-laden shales to sandstones harbouring sulphate-reducing

bacteria thus yielded an electron donor–acceptor couple required to stimulate sulphate reduction. Sulphate-reducing bacteria were thus apparently much less numerous in shales, but abundant and electron-donor-limited in sandstones.

The known interdependence of fermenters and sulphate-reducing bacteria suggested that a relatively complex microbial community was present. Organisms with other metabolic capabilities (such as methanogens and acetogens) could also be important in this system. No methane production could be measured in the sandstone–shale slurries, but acetate was produced in all incubations (Fig. 3), with the amount of acetate being far greater in sandstone slurries amended with shale than in incubations receiving only sandstone or shale. Pure cultures of both acetogenic and sulphate-reducing bacteria were also isolated from the rock material on the basis of their ability to grow by using hydrogen as the electron donor (L.R.K., unpublished results), providing further evidence for the existence of these processes in the sandstone units.

These results suggest an explanation for the presence of spatially discrete microbial ecosystems at sandstone–shale interfaces which are fuelled by organic matter deposited during the Cretaceous period. A similar model has been proposed for unconsolidated clay/sand aquifers of the late Cretaceous period^{14–16}. Our results build on this previous work, but in addition demonstrate that organic material can diffuse from shales to sandstones in consolidated rock, creating isolated microbial communities that exist in a narrow band near this interface. Such microbial communities are probably not unique to this Cretaceous lithologic system: they may exist near many other formations where biodegradation of organic carbon was constrained following deposition, and where adjacent higher-porosity sediments can help migration of microbial energy sources (see ref. 17, for example). Successional events could then select for the metabolic capacity to use the nutrients diffusing from adjacent sediment layers. Our results have far-reaching implications, in that they suggest that there are large areas of undiscovered biomass growing on ancient nutrient-rich deposits within isolated subsurface sedimentary rocks. They also explain the long-term survival of microbial communities in the subsurface and justify searching for microbial life in unexplored environments. □

Methods

Radioisotope imaging of sulphate-reduction activity. All manipulations were done in an anaerobic glove box under a nitrogen:hydrogen (5% H₂) headspace. Rock cores were fractured to reveal a new face, and ³⁵SO₄²⁻ (10 µCi per core) was applied to the core face by dripping 100 µl of 100 µM Na₂SO₄ solution on to it. After pretreatment of silver foils (8 cm × 8 cm) by washing consecutively with ethanol, acetone and hexane, the surface was oxidized in concentrated nitric acid. Foils were then washed in water, rewashed with the solvents, dried and autoclaved. The foil was then placed on the core face; the rock core was reassembled along fracture lines, clamped in position and incubated for 6 weeks. The foil was then washed with water to remove sulphate and the labelled sulphide adhering to the foil surface was analysed by radio-image analysis (Packard Instruments).

Analyses. To measure sulphate reduction, serum bottles were incubated with 5 ml filter-sterilized groundwater (obtained from the Cubero sandstone formation and amended with 10 µCi Na₂³⁵SO₄, resazurin (1 mg l⁻¹), NH₄Cl (20 mg l⁻¹) and K₂HPO₄ (10 mg l⁻¹) and adjusted to pH 7.6 by the addition of CO₂ to the headspace) and 0.75 g of ground sediments from 213 m depth or 1.5 g of sediments from 247 m. Original sulphate in these incubations was ~3.0 mM. To be certain that contamination of sediments did not occur from the surface of cores that had been exposed to drilling fluids, sediments were ground by drilling from the centre of the newly exposed face of the core with a sterile masonry drill bit; in this way, the grinding process did not pick up any rock from near the surface of the core. This experiment was repeated with sediments ground using a mortar and pestle rather than a drill bit, with similar results (data not shown). No activity was detected in autoclaved controls. Fluorescent microspheres were used as tracers during the drilling of these samples. No contamination of the 213-m sample could be detected. Slurry

subsamples (0.5 ml) were periodically removed and Cr(II)-reducible sulphur (sulphides, pyrite and sulphur) were measured using a modification of the procedure of Fossing and Jørgensen¹⁸. The sample was added directly to a stoppered 120-ml serum bottle containing 5 ml 1 M Cr(II)Cl₂ and 2.5 ml concentrated HCl. A 5-ml test tube containing 2.5 ml 10% zinc acetate was also in the bottle to act as a trap for volatilized sulphide. After a 3-day incubation, the trap was removed and the radioactivity counted. This technique traps sulphur from sulphides and pyrite with almost 100% efficiency¹⁹. Acetate was determined using a Dionex ion chromatograph (Dionex Instruments, CA). TOC was determined by difference from measurements of total and carbonate carbon. Total carbon was determined by combustion using a Leco carbon analyser, and carbonate carbon by digestion and titration (Method ASTM 513 Section G).

Received 7 October 1996; accepted 9 January 1997.

1. Pederson, K. *Earth-Sci. Rev.* **34**, 243–260 (1993).
2. Beeman, R. E. & Suflika, J. M. *Microb. Ecol.* **14**, 39–54 (1987).
3. Balkwill, D. L., Fredrickson, J. K. & Thomas, J. M. *Appl. Environ. Microbiol.* **55**, 1058–1065 (1989).
4. Amy, P. S., Haldeman, D. H., Ringelberg, D. B., Hall, D. H. & Russell, C. *Appl. Environ. Microbiol.* **58**, 3367–3373 (1992).
5. Lipman, C. B. *Science* **68**, 272–273 (1928).
6. Turner, H. G. *Science* **76**, 121–122 (1932).
7. Farrell, M. A. & Turner, H. G. *J. Bacteriol.* **23**, 155–162 (1932).
8. Fredrickson, J. K. & Phelps, T. J. in *Manual of Environmental Microbiology* (eds Knudsen, G., Stetzenbach, L., McInerney, M. M. & Walter, M.) 526–540 (Am. Soc. Microbiol., Washington, 1996).
9. Fredrickson, J. K. *et al. Geomicrobiol. J.* (in the press).
10. Kauffman, E. B. in *Treatise on Invertebrate Paleontology A* (eds Robinson, R. A. & Teichert, C.) 418–487 (Geol. Soc. Am., Boulder and Univ. of Kansas, Lawrence, 1979).
11. Pegram, P. thesis, New Mexico Inst. Mining and Technol. (1995).
12. Chambers, L. A. & Trudinger, P. A. *Geomicrobiol. J.* **1**, 249–293 (1979).
13. Sharma, P. K. & McInerney, M. J. *Appl. Environ. Microbiol.* **60**, 1481–1486 (1994).
14. McMahon, P. B. & Chapelle, F. H. *Nature* **349**, 233–235 (1991).
15. McMahon, P. B., Chapelle, F. H., Falls, W. F. & Bradley, P. M. *J. Sedim. Petrol.* **62**, 1–10 (1992).
16. Chapelle, F. H. & Lovley, D. R. *Appl. Environ. Microbiol.* **56**, 1865–1874 (1990).
17. McKinley, J. P. *et al. Geomicrobiol. J.* **14**, 23–39 (1997).
18. Fossing, H. & Jørgensen, B. B. *Biogeochemistry* **8**, 205–222 (1989).
19. Ulrich, G. A., Krumholz, L. R. & Suflika, J. M. *Appl. Environ. Microbiol.* **63** (in the press).

Acknowledgements. We thank Y. Cohen for developing the silver foil technique, P. Long, J. Fredrickson and T. Dewers for discussion, and F. Wobber for his continued support. This work was supported by the US Department of Energy, Subsurface Science Program.

Correspondence and requests for materials should be addressed to L.R.K. (e-mail: krumholz@ou.edu).

Visual decomposition of colour through motion extrapolation

Romi Nijhawan

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853, USA

The perception of yellow has played a central role in distinguishing two main theories of colour vision. Hering proposed that yellow results from the activation of a distinct retinal–neural mechanism, whereas according to the Young–Helmholtz–Maxwell view, yellow results from the combined activation of red and green cone mechanisms¹. When red and green images are presented separately to corresponding retinal locations in the two eyes, the resulting sensation is yellow^{1,2}. As the pathways from the two eyes do not converge until the cortex, this suggests that yellow can indeed arise from the central combining of separate red and green channels³. I now show that the reverse process can also occur; the visual system can decompose a ‘yellow’ stimulus into its constituent red and green components. A ‘yellow’ stimulus was created by optically superimposing a flashed red line onto a moving green bar. If the bar is visible only briefly, the flashed line appears yellow. If the trajectory of the green bar is exposed for sufficient time, however, the line is incorrectly perceived to trail the bar, and appears red. Motion processing occurs in the cortex rather than the retina in primates, and so the ability of motion cues to affect the perception of colour is consistent with the Young–Helmholtz–Maxwell notion of a ‘central synthesis’ of yellow.

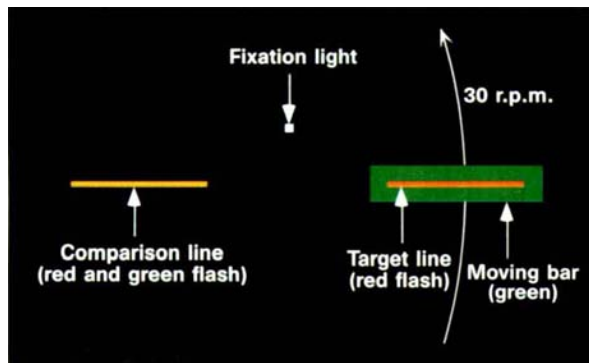


Figure 1 The physical display. Background luminance was 0.00 cd m^{-2} . A comparison line (red + green components) and a target line (red) were simultaneously flashed for 2 ms. The intensity of the comparison line's green component and the intensity of the green bar could be varied. The red target line and the red component of the comparison line were identical and invariant. Arbitrary intensities of the green bar and the comparison line's green component are shown. All coloured stimuli were produced with broadband filters. The green filter had zero transmittance cut-offs at 492 and 580 nm (peak 528 nm); the red filter had a cut-off at 592 nm (peak > 660 nm). For further filter specifications, see ref. 29. The fixation light was located 0.48° above the midpoint between the lines.

It is predominantly believed that colour and motion are processed by independent neural channels^{3–6}, but there are findings that the motion system can be stimulated by borders defined solely by chromatic contrast⁷. Humans can use colour to reliably distinguish the movement direction of very briefly exposed stimuli, further supporting the view that colour can independently contribute to motion perception⁸. The following display describes a previously unknown interaction between the mechanisms of colour and motion.

A green bar moved on a circular path on a dark background, while two lines, a comparison and a target, were flashed simultaneously for 2 ms. The target line was red and its flash was superimposed on the green bar (Fig. 1). Observers changed the colour of the comparison line to match that of the target line. The comparison line had a red component that was identical to the target line, and a green component that could be adjusted by the observer from zero intensity (resulting in a red comparison line) to maximum (resulting in a green–yellow comparison line). A green and a red filter (transmittance peaks 528 nm and > 660 nm, respectively) were used to produce the coloured stimuli.

Initially, the green bar was exposed at the location of the red target line so briefly that it appeared motionless. The comparison line was fixed at the green–yellow colour (maximum intensity of the green component) in this 'brief exposure' condition, and the bar intensity was changed until observers indicated a colour match between the lines (Fig. 2a). This bar intensity, determined for each observer, was next used in the 'extended exposure' condition in which the full trajectory of the bar was exposed (Fig. 2b). Observers now adjusted the colour of the comparison line to match that of the target line. Were the observers' matches based on the spectral composition of the target line, the comparison line's green component should have been set to the same value in the brief and extended exposure conditions. In other words, the green components of the comparison lines in Fig. 2a, b should have been set to the same (maximum) value. However, in the extended exposure condition the comparison line's green component was set to one fifth its maximum value. Thus, despite equal quantal absorption in the cones, the colour of the target line was perceived as green–yellow with brief exposure, and as reddish with extended exposure of the bar. Figure 3 shows data for three observers. In addition, observers reported on the perceived location of the target line relative to the bar. Consistent

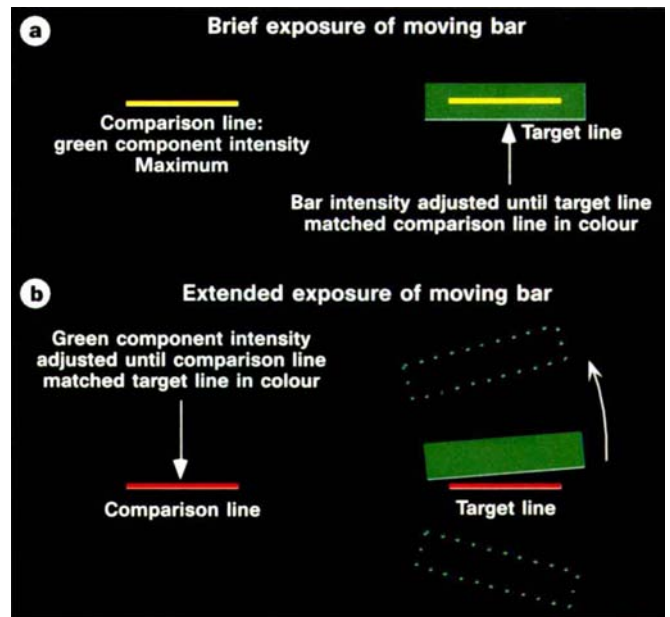


Figure 2 The brief- and extended-exposure conditions. **a**, In the brief-exposure condition, the comparison line's green component was fixed at its maximum intensity. The intensity of the green bar was changed until the observer reported that the colour of the target line matched that of the comparison line. With brief exposure, observers perceived the target line to be superimposed on the bar and its colour to be yellow. **b**, In the extended-exposure condition, the intensity of the green bar was fixed. Observers changed the intensity of the comparison line's green component until the colour of the comparison line matched that of the target line. With extended exposure, the target line appeared on the dark background, spatially lagging behind the green bar, and its colour appeared red. Note that in the experiment the spectral composition of the target line, and the target line's physical position relative to the bar, were identical in **a** and **b**.

with previous findings^{9,10}, the target line appeared to be superimposed on the bar in the brief-exposure condition (Fig. 2a), and on the background, trailing the bar, in the extended-exposure condition (Fig. 2b).

Could the bar's perceived motion *per se* have caused the target line to appear red, or is the perceived separation between the target line and the bar crucial? The smaller dimension of the bar was scaled up by a factor of ten so that the target line would appear superimposed on the bar, even while the bar was seen in motion. The colour matches obtained in this 'wide-bar' condition show only a slight tendency of the target line to appear red (Fig. 3), which suggests that the significant factor is the apparent non-overlap of the line and the bar. The wide-bar condition also eliminates the potential factors of chromatic adaptation¹¹ such as stray light from the green bar producing adaptive effects that dominate its colour mixing effects¹², and unmeasured eye-drifts causing prolonged retinal stimulation by green light.

The primate visual system is characterized by significant transit delays of neural signals between the photoreceptors and higher cortical areas. As a result, a moving object's retinal image location should lead the object's neural representation in any retinotopically organized cortical map. However, it has been suggested that an 'early' visual operation, mathematically akin to extrapolation, corrects for cortical lag and restores position correspondence between different processing levels for predictably moving objects^{10,13}. There is evidence suggesting that a mechanism for motion extrapolation might be present even in other animals, such as frogs¹⁴. As a result of motion extrapolation, the retinotopic site in the cortex maximally activated by a moving object at any given instant is the same as would be activated by a stationary object

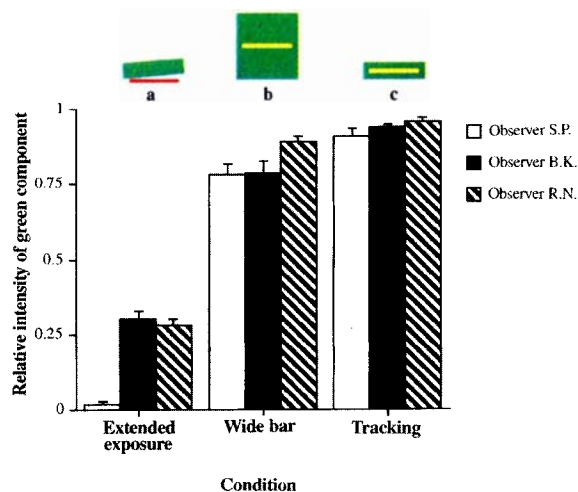


Figure 3 Three observers, two experienced (B.K. and R.N.) and one naive (S.P.), participated. The average bar intensity that produced a colour match between the lines in the brief-exposure condition for S.P., B.K. and R.N. was 1.41, 1.91 and 1.25 cd m^{-2} , respectively. The maximum intensity of the comparison line's green component, which caused the line to appear green-yellow, is represented as '1' on the ordinate. The average intensity of the green component set by the observers, in the extended-exposure, wide-bar, and tracking conditions, is reported as a proportion between 0 and 1. In the 'extended-exposure' condition (motion visible) the value of the comparison line's green component set by the observers was, on average, one fifth its maximum value. The 'wide-bar' condition (motion visible) tested the role of motion *per se* in causing the red appearance of the target line. The 'tracking' condition, which restored alignment between the moving green bar and the flashed target line, tested the role of spatial separation. In both the 'wide-bar' and 'tracking' conditions the green bar was sensed as moving, but the target line appeared to overlap the bar. These conditions produced only a small shift in the appearance of the target line towards red, suggesting that spatial separation was the crucial factor. Inset **a**, the observers' percept in the 'extended-exposure' condition—the target line appeared red and not to overlap the green bar. Inset **b**, the percept in the 'wide-bar' condition—the target line appeared yellow and to overlap the bar. Inset **c**, the percept in the 'brief-exposure' and 'tracking' conditions—the target line appeared yellow and to overlap the bar in both.

located where the moving object is at that instant. In contrast, a brief flash is unpredictable and the nerve impulses triggered by the flash arrive in the cortex after the expected delay. Thus, at a given time the flash and the moving object stimulate separate retinotopic locations in the cortex, with the moving object in a leading position. This is the basis of the observers' perception that the moving bar leads the flashed target line.

This extrapolation and delay hypothesis has further support in that smooth pursuit of the moving item restores its alignment with the flashed item (R.N., manuscript in preparation). The retinal displacement of a flashed object, while the eyes are tracking, will be insignificant—0.8- μs flashes have produced comparable results. It is known that eye position and perceived location of retinally stable objects, like ordinary after images, are perfectly correlated. What about flashed objects? The extrapolation and delay hypothesis assumes identical processing of flashed and moving items during smooth pursuit and fixation. However, eyes in pursuit can rotate significantly in the time it would take photoisomerizations triggered by the flash to culminate in cortical signals¹⁵. It is suggested that, like after images, flashed objects are also seen in the direction in which the eyes point. As the eyes, by definition, always point towards the tracked object, the flash appears to overlap the moving object. A further prediction of this hypothesis, that a flashed object should appear displaced in the direction of pursuit relative to a stationary target, has been experimentally verified.

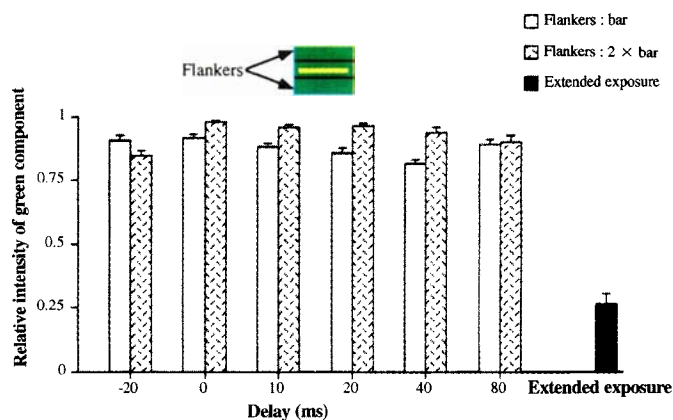


Figure 4 Abscissa, time delays between the flash of the lines and the flash of the flanker masking bars. A negative delay indicates that the flankers were flashed before the lines. Several delays were used to ensure that at least one delay would significantly reduce the original bar's visibility. The observer, as before, matched the colour of the two lines. The height of the data bars corresponds to the intensity of the comparison line's green component that produced a colour match for observer R.N. The right-most data bar corresponds to the extended exposure condition with no flankers. Relative to the extended exposure condition, the other data bars show a relatively small decrease in the intensity of the comparison line's green component. Other observers gave similar results. Each flanking bar in the inset is as wide as the original bar.

The perceived overlap of the flashed target line and the moving green bar obtained with pursuit eye movements was used to test whether the apparent separation between these items is necessary for the colour effect. In the 'tracking' condition, observers were instructed to smoothly pursue the bar and report the location and the colour of the target line. Observers reported the target line to be centred precisely on the bar, and its colour to be yellow (Fig. 3). This outcome further reinforces my finding that the apparent separation of the bar and the target line is critical for the target line's red appearance, and that motion *per se* is not sufficient. The target line's red appearance in the extended-exposure condition may be an instance of the phenomenon of colour constancy^{16,17} in which the bar's antecedent colour signals are used to discount its spectral contribution. However, this would predict a similar outcome in the wide-bar and tracking conditions as well, which is not observed.

Flashed and moving items differ in another important way. The visibility of a flashed stimulus persists beyond the physical flash by about 120 ms (refs 18, 19), whereas the 'smear' due to the gradual decay of neural activity left behind by a moving object is not visible to observers. The mechanisms specialized to detect motion also reduce motion smear^{20,21}. The green bar signals motion with extended exposure but not with brief exposure. Arguably, the green bar's prolonged visibility in the brief-exposure condition could bias the perceived colour of the target line towards green. Thus, a persisting green bar of lower intensity might produce the same colour shift in the target line as a green bar of higher intensity that does not persist. As the bar intensity in the brief exposure (Fig. 2a) and extended exposure (Fig. 2b) conditions is the same, the colour of the target line could be biased towards red in the latter condition. If this hypothesis is correct, then a similar bias should be observed if the visibility of the bar is reduced by a visual mask²² in place of motion. A mask consisting of two green bars, presented with a delay of ~40 ms after a target bar, significantly reduces the target bar's visibility^{23,24}. The brief exposure condition was modified so that masking bars were presented at several delays to ensure that

the original bar's visibility would be sufficiently reduced for at least one of these delays²⁵ (Fig. 4). Although the mask was effective, it produced only a slight shift in the target line's colour towards red (Fig. 4), so failing to support the persistence explanation.

Further experiments will reveal the neural basis of this colour phenomenon, which cannot have a retinal origin as the primate retina is unable to distinguish the direction of movement of a target²⁶. The effect is probably a by-product of motion extrapolation, which could be accomplished by cortical neurons that demonstrate 'predictive remapping'. For these neurons, the response latency for a target entering the cell's receptive field from some other location is lower than the latency for a target with an abrupt onset in that cell's receptive field²⁷. At a given instant, two sets of active cells of this type, one stimulated by a moving object and the other by a retinally superimposed flash, will thus code non-overlapping retinotopic locations. Given the reported interaction between colour and motion, I suggest that some cells that demonstrate predictive remapping are also stimulated by colour input. Finally, my findings not only support the Young-Helmholtz-Maxwell view of colour vision, but also suggest that retinally overlaid red and green signals yield the perception of yellow if, and only if, these signals originate from the same retinotopic locations in the cortex. □

Methods

A d.c. motor connected to a speed controller revolved a green bar ($1.59^\circ \times 0.20^\circ$) anti-clockwise on a circular path (diameter, 7.63°) at 30 r.p.m. (instantaneous speed, $11.46^\circ \text{ s}^{-1}$). Two horizontal lines (length, 1.27°), separated by 0.91° , were flashed simultaneously for 2 ms at 0.5 hertz. A mirror-type beam-splitter was used to present the flashed lines in the optical plane of the bar. The flash was synchronized with the bar so that the target line was superimposed on the bar. Observers used a chin rest to view the display binocularly through natural pupils. Ambient room lights were turned on about once every 10 s for a comparable period in order to maintain the observer's light adaptation level and pupil size.

Brief-exposure condition. The bar was visible only through a $1.59^\circ \times 0.20^\circ$ window. The duration of visibility was ~ 17.45 ms. The colour of the comparison line was held constant at green-yellow. On ten trials the experimenter adjusted the bar's intensity until the observer indicated that the target line matched the comparison line in colour. The average of the bar intensities for each observer was used in all conditions tested.

Extended-exposure condition. The complete trajectory of the green bar was visible. The bar's intensity was set at a value that produced a colour match of the lines in the brief-exposure condition. On ten trials the observer adjusted the intensity of the comparison line's green component to achieve a colour match between the lines.

Wide-bar and tracking conditions. A wide bar ($1.59^\circ \times 2.0^\circ$) was used to test the contribution of motion *per se* to the colour effect (see text). In the tracking condition, observers used smooth-pursuit eye movements to follow the bar of original size ($1.59^\circ \times 0.20^\circ$).

Persistence reduction through visual masking. Introduction of a visual mask in the brief exposure condition required an additional viewing channel which involved several changes in the optical arrangement. Consequently, the brief and extended exposure conditions were repeated. The mask consisted of two flanking bars, abutting the original bar with no gap or overlap, flashed for 10 ms. Each flanking bar was either $1.59^\circ \times 0.20^\circ$ or $1.59^\circ \times 0.40^\circ$. The intensity of the flanking bars was equated to that of the original bar by using modified 'minimally distinct border' technique²⁸.

Received 7 May 1996; accepted 9 January 1997.

- Hurvich, L. M. & Jameson, D. The binocular fusion of yellow in relation to color theories. *Science* **114**, 199–202 (1951).
- Hecht, S. On the binocular fusion of colors and its relation to theories of color vision. *Proc. Natl Acad. Sci. USA* **14**, 237–241 (1928).
- Zeki, S. M. The functional organization of projections from striate to prestriate visual cortex in the rhesus monkey. *Cold Spring Harb. Symp. Quant. Biol.* **40**, 591–600 (1975).
- Livingstone, M. S. & Hubel, D. H. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *J. Neurosci.* **7**, 3416–3468 (1987).
- Ramachandran, V. S. & Gregory, R. L. Does colour provide an input to human motion perception? *Nature* **275**, 55–56 (1978).

- Cavanagh, P., Tyler, C. W. & Favreau, O. E. Perceived velocity of moving chromatic gratings. *J. Opt. Soc. Am.* **1**, 893–899 (1984).
- Saito, H., Tanaka, K., Isono, H., Yasuda, M. & Mikami, A. Directionally selective response of cells in the middle temporal area (MT) of the macaque monkey to the movement of equiluminous opponent color stimuli. *Exp. Brain Res.* **75**, 1–14 (1989).
- Cropper, S. J. & Derrington, A. M. Rapid colour-specific detection of motion in human vision. *Nature* **379**, 72–74 (1996).
- MacKay, D. M. Perceptual stability of a stroboscopically lit visual field containing self-luminous objects. *Nature* **181**, 507–508 (1958).
- Nijhawan, R. Motion extrapolation in catching. *Nature* **370**, 256–257 (1994).
- Hurvich, L. M. *Color Vision* (Sinauer, Sunderland, Massachusetts, 1981).
- Walraven, J. Spatial characteristics of chromatic induction; the segregation of lateral effects from straylight artefacts. *Vision Res.* **13**, 1739–1753 (1973).
- Khurana, B. & Nijhawan, R. Extrapolation or attention shift? *Nature* **378**, 565–566 (1995).
- Aho, A.-C., Donner, K., Helenius, S., Olesen Larsen, L. & Reuter, T. Visual performance of the toad (*Bufo bufo*) at low light levels: retinal ganglion cell responses and prey-catching accuracy. *J. Comp. Physiol.* **172**, 671–682 (1993).
- Richards, W. Visual suppression during passive eye movement. *J. Opt. Soc. Am.* **58**, 1159–1160 (1968).
- Land, E. H. & McCann, J. J. Lightness and the retinex theory. *J. Opt. Soc. Am.* **61**, 1–11 (1971).
- Gilchrist, A. The perception of surface blacks and whites. *Sci. Am.* **24**, 88–97 (1979).
- Efron, R. The minimum duration of a perception. *Neuropsychologia* **8**, 57–63 (1970).
- Hogben, J. H. & Di Lollo, V. Perceptual integration and perceptual segregation of brief visual stimuli. *Vision Res.* **14**, 1059–1069 (1974).
- Burr, D. Motion smear. *Nature* **284**, 164–165 (1980).
- Morgan, M. J. Perception of continuity in stroboscopic motion: A temporal frequency analysis. *Vision Res.* **19**, 491–500 (1979).
- Kahneman, D. Method, findings, and theory in studies of visual masking. *Psychol. Bull.* **70**, 404–425 (1968).
- Alpern, M. & Rushton, W. A. H. The specificity of the cone interaction in the after-flash effect. *J. Physiol.* **176**, 473–482 (1965).
- Yellott, J. I. & Wandell, B. A. Color properties of the contrast flash effect: Monoptic vs dichoptic comparisons. *Vision Res.* **16**, 1275–1280 (1976).
- Di Lollo, V. & Hogben, J. H. Suppression of visible persistence as a function of spatial separation between inducing stimuli. *Percept. Psychophys.* **41**, 345–354 (1987).
- Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (Lond.)* **195**, 215–243 (1968).
- Duhamel, J.-R., Colby, C. L. & Goldberg, M. E. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**, 90–92 (1992).
- Boynton, R. M. in *Visual Psychophysics and Physiology* (eds Armington, J. C., Krauskopf, J. & Wooten, B. R.) 193–207 (Academic, New York, 1978).
- Zeki, S. The representation of colours in the cerebral cortex. *Nature* **284**, 412–418 (1980).

Acknowledgements. This work would not have been possible without the constant interest and support of B. Khurana. I thank P. Cavanagh for his contributions; K. Nakayama for the use of the Vision Sciences Laboratory, Harvard University; T. Reuter and the A. Treisman group for their interest; A. Derrington, R. Singh, I. Thornton, A. Gilchrist, J. Cutting, M. Morgan and O. Braddick for comments and discussion; and E. Snyder and F. Horan for technical assistance.

Correspondence and requests for materials should be addressed to the author (e-mail: rn15@cornell.edu).

Adaptation of retinal processing to image contrast and spatial scale

Stelios M. Smirnakis[†], Michael J. Berry[‡], David K. Warland[‡], William Bialek[§] & Markus Meister[‡]

Departments of ^{*} Physics, and [‡] Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA
[†] Division of Health Sciences and Technology, Harvard Medical School, 260 Longwood Avenue, Boston, Massachusetts 02115, USA
[§] NEC Research Institute, Princeton, New Jersey 08540, USA

Owing to the limited dynamic range of a neuron's output, neural circuits are faced with a trade-off between encoding the full range of their inputs and resolving gradations among those inputs. For example, the ambient light level varies daily over more than nine orders of magnitude¹, whereas the firing rate of optic nerve fibres spans less than two². This discrepancy is alleviated by light adaptation³: as the mean intensity increases, the retina becomes proportionately less sensitive. However, image statistics other than the mean intensity also vary drastically during routine visual processing. Theory predicts that an efficient visual encoder should adapt its strategy not only to the mean, but to the full shape of the intensity distribution^{4–6}. Here we report that retinal ganglion cells, the output neurons of the retina, adapt to both image contrast—the range of light intensities—and to spatial correlations within the scene, even at constant mean intensity. The adaptation occurs on a scale of seconds, one hundred times

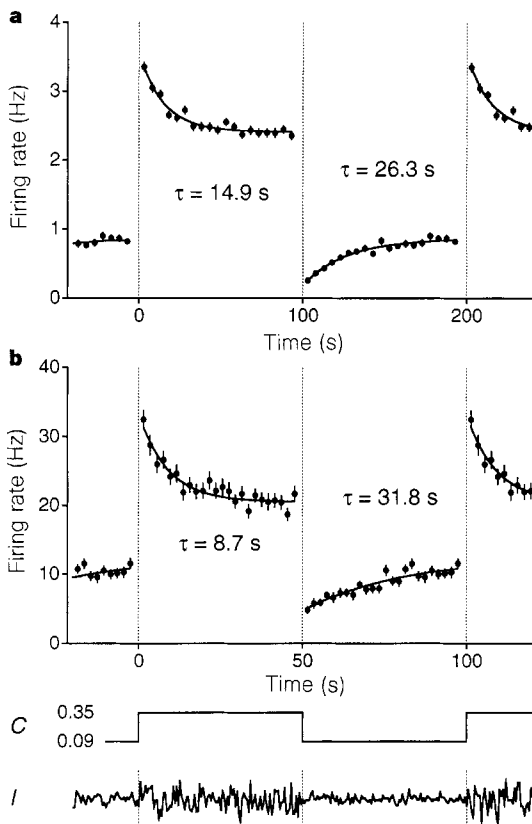


Figure 1 Firing rate of a salamander (a) and a rabbit (b) ganglion cell under spatially uniform flicker stimulation, alternating every 100 (a) or 50 (b) seconds between contrast values of 0.09 and 0.35. Average firing rate values were computed in 5-s (a) or 2-s (b) time bins over 100 (a) or 25 trials (b). Continuous lines are exponential fits with decay time τ . The first and last segments are periodic repeats of the data. Trace below (b) shows the time course of the contrast, C . Bottom trace illustrates the time course of the flickering intensity, I ; note that the random flicker sequence was different in each stimulus trial. Similar gradual changes in the firing rate were unambiguous in 75% of salamander and 51% of rabbit cells. Another 18% in salamander and 31% in rabbit showed the gradual firing rate decline after a contrast increase but no detectable recovery following a contrast decrease.

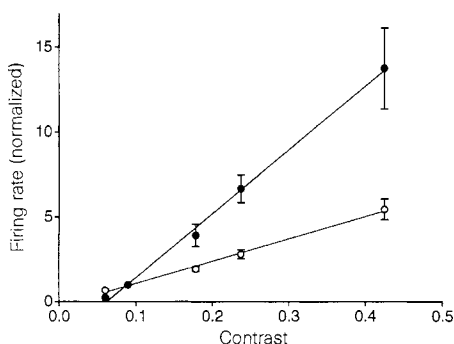


Figure 2 Initial firing rate (R_i , closed symbols) and final firing rate (R_f , open symbols) following a contrast step from $C = 0.09$ to values ranging between 0.06 and 0.43. For each cell, R_i and R_f were calculated as in Table 1 and normalized to yield $R_i = 1$ at $C = 0.09$, then averaged over 53 salamander ganglion cells from 2 retinæ. Lines are linear fits.

more slowly than the immediate light response, and involves 2–5-fold changes in the firing rate. It is mediated within the retinal network: two independent sites of modulation after the photoreceptor cells appear to be involved. Our results demonstrate a remarkable plasticity in retinal processing that may contribute to the contrast adaptation of human vision⁷.

We recorded the spike trains of ganglion cells in the isolated retina of a tiger salamander or rabbit. The photoreceptor layer was presented with random flicker stimuli whose mean intensity remained constant throughout, but whose second-order statistics changed abruptly at long intervals (see Methods). The goal was to monitor changes in retinal processing that are brought about by such a switch in the properties of the artificial visual environment. The first set of experiments employed a spatially uniform flickering field; its mean intensity was kept constant, but the width of the intensity distribution increased or decreased abruptly. Following such a step increase in contrast, the firing rate of most retinal ganglion cells increased abruptly, then decayed exponentially to a much lower value (Fig. 1 and Table 1). Following the opposite contrast step, the firing rate dropped abruptly and then recovered exponentially. These gradual changes in the firing rate suggest an adaptive modification of ganglion cell response properties following a step in contrast. For most salamander cells, the decay of the firing rate after a contrast increase was 2 to 3 times faster than its recovery following a contrast decrease (Table 1). Rabbit ganglion cells adapted very similarly (Fig. 1b and Table 1), even though their absolute firing rates were an order of magnitude higher than those of salamander ganglion cells. 'On' and 'off' cells showed no qualitative differences in their adaptation behaviour to this stimulus.

The initial firing rate immediately following a contrast jump

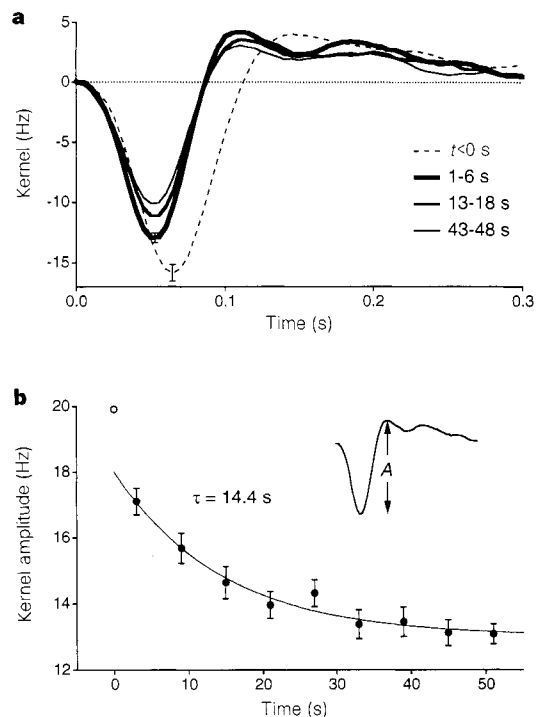


Figure 3 a, The effect of a contrast step from $C = 0.09$ to 0.35 on the linear kernel of a salamander ganglion cell (from Fig. 1a): just before the step (broken line), after the step at 1–6 s (thick line), 13–18 s (medium line), and 43–48 s (thin line). The kernel $k(t)$ was computed with 5-ms bins over the respective time interval and averaged over 50 stimulus trials (see Methods). **b**, Peak-to-peak amplitude (inset) of the kernel from the cell in (a), as a function of time relative to the contrast transition. Open symbol represents value at $t < 0$. Continuous line is an exponential fit to data after the step, with decay time τ . A similar decrease of the kernel amplitude was observed in 93% of salamander and 85% of rabbit cells.

increased linearly with the final contrast (Fig. 2). The steady-state firing rate, after the exponential transient, also varied linearly with the contrast. However, the latter curve was less than half as steep as the former, a clear sign of contrast adaptation. Using the same criterion, but different stimuli, a small but significant contrast adaptation effect has been demonstrated in the cat's lateral geniculate nucleus⁸.

To determine how each ganglion cell responded to the flickering light, we computed the linear kernel of its firing rate with respect to the stimulus intensity^{9,10}; this can also be viewed as the linear response to a brief flash of light (see Methods). A step increase in contrast triggered a rapid change in the shape of this kernel towards a faster waveform (Fig. 3a, broken and thick lines). This was followed by a gradual decrease of the kernel amplitude (Fig. 3 and Table 1). Following the opposite contrast step, the kernel shape reverted and its amplitude gradually increased. The fact that the slow changes in kernel amplitude and firing rate occur with a very similar time course (Figs 1a and 3b) suggests that the rate modulations truly reflect a change in the sensitivity of the light response. At low contrast, a high sensitivity serves to represent effectively the narrow range of intensities. Following a sudden transition to high contrast, the sensitivity gradually decreases to encode the broader intensity range.

Previous studies of retinal signalling in cat¹¹ and catfish⁹ indicated that ganglion cell response properties change as a function of ambient contrast. This 'contrast gain control' occurred within tens of milliseconds¹² of a contrast step, which is less than the photoreceptor integration time. Figure 3a also shows a shift of ganglion cell sensitivity towards higher temporal frequencies that occurs immediately after the contrast step, as seen in cat¹¹. These fast changes can be viewed as a nonlinear feature of the instantaneous

light response¹³, rather than as gradual adaptation to scene statistics. By comparison, the subsequent adaptation phase documented here requires seconds to tens of seconds. This is similar to the time in which an animal might move into an environment of substantially different texture, or in which the scene illumination might change from direct to diffuse light, producing a change in contrast.

Some neural element within the retina must sense the contrast and control the change in sensitivity. To estimate the size of this neuron's receptive field, we stimulated the salamander retina alternately with a flickering checkerboard of varying square size, D , and spatially uniform flicker of the same mean and contrast (see Methods). Under these conditions, only neurons whose receptive field extended over more than one square of the checkerboard experienced a change in the statistics of their synaptic input as a result of the transition to uniform flicker.

Figure 4a illustrates the firing rate of a salamander ganglion cell in such an experiment. Following a transition from uniform to checkerboard stimulation, the firing rate suddenly increased, then gradually declined over several seconds to a steady state. Following the reverse transition, the firing rate suddenly decreased, then gradually recovered. Judging from the final steady-state firing rate, the neuron was driven more strongly by checkerboards than by uniform fields, consistent with the centre-surround antagonism in its receptive field. During the adaptation phase, the neuron's sensitivity always declined when exposed to a stronger stimulus, and recovered under a weaker stimulus, analogous to the observations with uniform stimulation at varying contrast (Fig. 1). Thus, the same mechanism may account for both observations (Fig. 5). Such behaviour was seen in all 'on' cells.

However, the great majority of 'off' ganglion cells behaved like the neuron shown in Fig. 4b. During the adaptation phase the neuron's

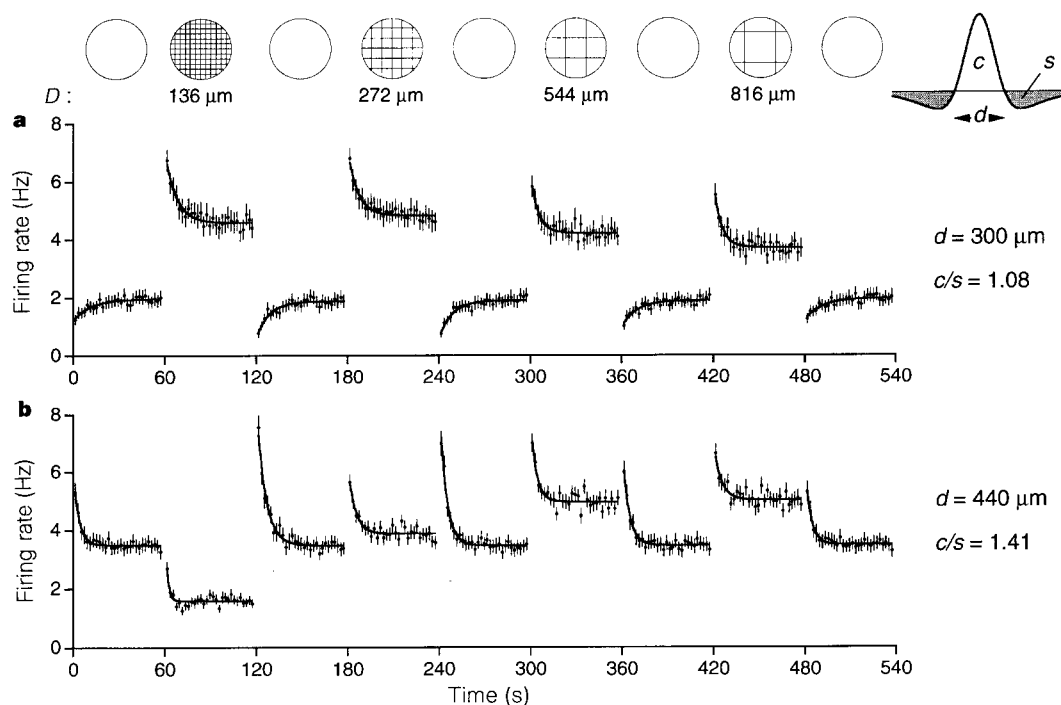


Figure 4 Adaptation to a change in spatial scale for two salamander ganglion cells, 'on' type (a) and 'off' type (b). Top row illustrates time course of the stimulus during a trial, alternating between uniform and checkerboard stimulation at varying square size, D . The checkerboard was also shifted every 30 ms by a spatial offset chosen at random on a 136- μ m grid; this ensured that checker boundaries would not consistently fall on the same neural receptive fields. Contrast remained at $C = 0.24$ throughout. Average firing rates were computed in 2-s intervals over 50 trials. Lines are exponential fits. The last data segment (480–540 s) is identical to the first (0–60 s). The receptive field profiles (see Methods) of

both cells are summarized (right) by the diameter of the centre region, d , and the ratio of integrated sensitivity in the positive centre to that in the negative surround, c/s . Note that the steady-state firing rates are maximal when $D \approx d$, and are suppressed more under uniform illumination when the surround is stronger, as in a. 12% of recorded neurons showed adaptation as in a, 82% as in b; for 6% the firing rate adapted downwards only following the transition from uniform to checkerboard stimulation. In the rabbit retina these stimuli produced qualitatively similar adaptation of the ganglion cell firing rate, but there was no sharp distinction between 'on' and 'off' cells.

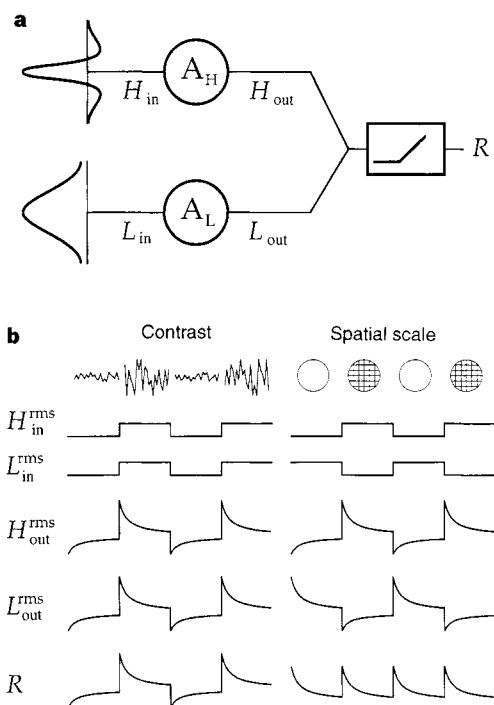


Figure 5 a, Model of contrast adaptation. Two pathways lead from the stimulus to the ganglion cell: one (H, 'high pass') integrates the stimulus with a spatially antagonistic receptive field; the other (L, 'low pass') has a broad receptive field. Each pathway contains an adaptive neuron (A_H , A_L) which gradually increases its sensitivity when its input signal (H_{in} , L_{in}) varies over a small range, and gradually decreases its sensitivity when its input range is large. The output signals of the two pathways (H_{out} , L_{out}) are pooled and rectified to produce a positive ganglion cell firing rate (R). **b**, Predictions for the firing rate under variations in contrast (left) or checkerboard-square size (right). The range of signals in the H and the L pathways are indicated by their root-mean-square variations, H^{rms} and L^{rms} . When the flicker contrast is stepped up (left), both pathways are driven more strongly and thus adapt by decreasing their sensitivity, resulting in a gradual decline in the firing rate. The opposite is predicted when the contrast steps down. When switching from uniform to checkerboard stimulation (right), the H pathway is driven more strongly and the L pathway more weakly. Thus their sensitivities adapt in the opposite direction. As the downward adaptation transients are larger than the upward transients (Fig. 1), the net effect is a gradual decline of the firing rate. When switching back to uniform stimulation, the roles of H and L pathways are switched, but their pooled activity again results in a gradual decline of the firing rate. This can explain the behaviour of salamander 'off' cells in Fig. 1a (left) and Fig. 4b (right). 'On' cells behave as though they lack the L pathway, or at least its adaptive control, because their firing rate under both conditions (Figs 1a and 4a) varies as the output signal in the H pathway.

Table 1 Characteristics of adaptation transients induced by changes in contrast

	Contrast		τ (s)		R/R_i		A_i/A_f	
	C_{low}	C_{high}	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$
Salamander	0.06	0.18	9.1 ± 0.9 (44)	17.8 ± 1.7 (24)	2.48 ± 0.31 (44)	0.18 ± 0.04 (24)	1.48 ± 0.07 (42)	0.56 ± 0.06 (17)
	0.09	0.35	8.8 ± 1.1 (20)	28.5 ± 2.7 (17)	1.69 ± 0.14 (20)	0.37 ± 0.06 (17)	1.26 ± 0.05 (20)	0.60 ± 0.07 (16)
Rabbit	0.09	0.35	12.6 ± 1.28 (36)	27.3 ± 2.7 (16)	2.26 ± 0.27 (36)	0.27 ± 0.06 (16)	1.45 ± 0.13 (29)	0.68 ± 0.07 (16)

Changes in contrast from C_{low} to C_{high} are indicated by an upward arrow, and from C_{high} to C_{low} by a downward arrow. For each cell, the time course of the firing rate was fitted with an exponential of time constant τ (Fig. 1) and extrapolated to yield the initial and final firing rate, R_i and R_f . The kernel amplitude, A , was taken as the peak-to-peak amplitude of the waveform (Fig. 3b), and its initial and final values, A_i and A_f , were measured over the first and the last 5-s interval of the contrast segment respectively. The parameters τ , R/R_i , and A_i/A_f were averaged over all cells that showed adaptation (Figs 1 and 3), except when τ or A could not be determined accurately. The number of cells for each estimate is given in parentheses. Uncertainties denote standard error.

sensitivity was initially high and gradually declined following every stimulus transition, irrespective of whether the final rate increased or decreased. In fact, the final rate under 272- μ m squares was almost identical to that under uniform stimulation, yet the transition produced a large transient during which the rate changed by more than a factor of two. This behaviour cannot be explained by a single mechanism of local contrast adaptation, but could result from independent adaptation in two pathways, one of which responds best at high spatial frequencies, the other at low frequencies (Fig. 5).

In these experiments, both the mean and the local contrast were held constant, and the adaptation process was triggered purely by a change in the spatial scale of the scene. These results place strong constraints on the site of adaptation. In particular, it does not occur in photoreceptors: The receptive field of cones—including their electrical coupling which extends over $\sim 25 \mu$ m in turtle¹⁴—is too small to detect the difference between checkerboard and uniform stimulation. In the scheme proposed in Fig. 5, the H pathway must have a narrow centre, with a large surround extending over more than 800 μ m (the largest square size tested). Bipolar cells—with their broad antagonistic surrounds from the horizontal cell syncytium^{15,16}—could serve this role, as could the ganglion cell itself. The L pathway could be implemented by wide-field amacrine cells in the inner retina^{17,18}. These assignments and the scheme in Fig. 5 are tentative, but they provide one framework for explaining the contrast adaptation behaviour of ganglion cells.

The effects reported here share several characteristics with contrast adaptation observed in human subjects and in the responses of cortical neurons. After adapting to a drifting grating, the threshold

for detecting a similar test grating is raised about 5-fold⁷, comparable in magnitude to the observed changes in the firing rate of retinal ganglion cells (Table 1, and see ref. 19). The time constants of psychophysical contrast adaptation also range from second to tens of seconds^{7,20}. The loss of sensitivity after a contrast increase and the recovery after a contrast decrease often follow a different time course^{21–23}, a consistent feature in our recordings (Fig. 1 and Table 1; $P < 0.001$). Although this suggests that retinal processing is important in psychophysical contrast adaptation, cortical mechanisms certainly contribute as well, because there is partial interocular transfer of adaptation^{7,20,24} and neurons in the cat visual cortex adapt more strongly to grating stimuli than neurons in the lateral geniculate nucleus⁸.

In summary, it appears that visual processing in the retina is considerably more adaptive than previously acknowledged and adjusts not only to the mean illumination but also to both the range of intensity fluctuations and their spatial scale. Essentially identical behaviour was observed in a mammalian and an amphibian species, and related effects are seen in an insect visual system⁵, suggesting that this strategy is a general principle of retinal processing. More broadly, it might be expected that any neural circuit would benefit from an adaptive control that responds to changes in the statistics of its inputs. The retina, with its accessible and well studied circuitry, is ideal for studying the underlying mechanisms. □

Methods

Recording. Retinae of larval tiger salamanders and Dutch belted rabbits were isolated in darkness into oxygenated Ringer's medium¹⁰ (salamander) or Ames'

solution (rabbit; Sigma A-1420 supplemented with 22.6 mM Na_2HCO_3 , 4.4 mM D-glucose, pH 7.4; maintained at 37°C during recording). A piece of retina 2–4 mm on a side was placed onto a flat multi-electrode array for extracellular recording from the ganglion cell layer^{10,25}. Results are reported from 71 cells in 4 salamander retinas, 55 cells in 2 rabbit retinas (Figs 1, 2, 3; Table 1), and 33 cells in 2 salamander retinas (Fig. 4). In the salamander, 14% of recorded ganglion cells were 'on' and 86% 'off'. In the rabbit, 33% were 'on', 51% 'off', and 16% had mixed response properties.

Stimulation. Stimuli were projected from a computer monitor onto a 3.25-mm-diameter aperture on the retina. Patterns consisted of a single uniform field of flickering light (Figs 1–3) or a flickering checkerboard with independently modulated square fields (Fig. 4), ranging in size, D , from 0.068 to 0.816 mm. The light intensity of each field was chosen every $\Delta t = 30$ ms at random from a gaussian probability distribution with mean M and standard deviation W . Contrast, C , was defined as W/M . The field size D sets the spatial scale of the stimulus. All experiments used white light²⁶ with a mean intensity of $M = 4.2 \text{ mW m}^{-2}$ at the retina. The equivalent photon flux at the peak absorption wavelength (λ_{max}) was 5,680 photons $\mu\text{m}^{-2} \text{ s}^{-1}$ for the salamander's red cone photoreceptor ($\lambda_{\text{max}} = 630 \text{ nm}$), and 5,150 photons $\mu\text{m}^{-2} \text{ s}^{-1}$ for the rabbit's green cone ($\lambda_{\text{max}} = 523 \text{ nm}$)^{27,28}.

Analysis. In a typical experiment, we interleaved segments of two stimulus types in the order $A_1B_1A_2B_2A_3B_3\dots$, where the A_i represents segments with the same contrast and spatial scale, but different random flicker sequences, and B_i represent segments with another contrast or spatial scale. Individual segments of A and B lasted either 50 s or 100 s, and recordings extended over 25 to 100 AB trials. The mean firing rate at a given time around the A-to-B transition (Figs 1, 2, 4) was computed by counting the spikes in the corresponding short time bin of all the AB trials, and dividing by the number of trials and the length of the time bin. The first-order Wiener kernel (Fig. 3) of the response was computed by correlating the stimulus intensity $I(t)$ and the firing rate $R(t)$: $k(t) = \frac{1}{WC} \frac{1}{T} \int_0^T (I(t') - M)R(t' + t)dt'$. This represents the linear effect on the firing rate from a flash of light with integrated intensity $M \cdot \Delta t$. The kernel at a given time around the transition was computed by performing the integral only over the corresponding time bins of all individual trials. The spatial receptive field (Fig. 4) was determined from steady-state responses to checkerboard stimulation with 136- μm squares: the kernel $k(t)$ was computed for each square, and its peak amplitude plotted as a function of position. All error bars in figures represent the standard error across trials.

Received 10 September 1996; accepted 9 January 1997.

- Rushton, W. A. Visual Adaptation (The Ferrier Lecture, 1962). *Proc. R. Soc. Lond. B* **162**, 20–46 (1965).
- Barlow, H. B. Critical limiting factors in the design of the eye and visual cortex (The Ferrier Lecture, 1980). *Proc. R. Soc. Lond. B* **212**, 1–34 (1981).
- Shapley, R. & Enroth-Cugell, C. Visual Adaptation and Retinal Gain Controls. *Progr. Ret. Res.* **3**, 263–346 (1984).
- Laughlin, S. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch.* **36**, 910–912 (1981).
- de Ruyter van Steveninck, R. R., Bialek, W., Potters, M. & Carlson, R. H. Statistical adaptation and optimal estimation in movement computation by the blowfly visual system. *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics* 302–308 (1994).
- DeWeese, M. & Bialek, W. Information flow in sensory neurons. *II Nuovo Cimento* **17D**, 733–738 (1995).
- Blakemore, C. & Campbell, F. W. On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *J. Physiol.* **203**, 237–260 (1969).
- Ohzawa, I., Sclar, G. & Freeman, R. D. Contrast gain control in the cat's visual system. *J. Neurophysiol.* **54**, 651–667 (1985).
- Sakai, H. M. & Naka, K. Signal transmission in the catfish retina. V. Sensitivity and circuit. *J. Neurophysiol.* **58**, 1329–1350 (1987).
- Meister, M., Pine, J. & Baylor, D. A. Multi-neuronal signals from the retina: acquisition and analysis. *J. Neurosci. Meth.* **51**, 95–106 (1994).
- Shapley, R. M. & Victor, J. D. The effect of contrast on the transfer properties of cat retinal ganglion cells. *J. Physiol.* **285**, 275–298 (1978).
- Victor, J. D. The dynamics of the cat retinal X cell centre. *J. Physiol.* **386**, 219–246 (1987).
- Wang, J.-L. & Naka, K.-I. Contrast gain control in the lower vertebrate retinas. *Soc. Neurosci. Abstr.* **21**, 1644 (1995).
- Copenhagen, D. R. & Green, D. G. The absence of spread of adaptation between rod photoreceptors in turtle retina. *J. Physiol.* **369**, 161–181 (1985).
- Dacheux, R. F. & Raviola, E. Horizontal cells in the retina of the rabbit. *J. Neurosci.* **2**, 1486–1493 (1982).
- Hare, W. A. & Owen, W. G. Spatial organization of the bipolar cell's receptive field in the retina of the tiger salamander. *J. Physiol.* **421**, 223–245 (1990).
- Werblin, F., Maguire, G., Lukasiewicz, P., Eliasof, S. & Wu, S. M. Neural interactions mediating the detection of motion in the retina of the tiger salamander. *Visual Neurosci.* **1**, 317–329 (1988).
- Bloomfield, S. A. Relationship between receptive and dendritic field size of amacrine cells in the rabbit retina. *J. Neurophysiol.* **68**, 711–725 (1992).
- Barlow, H. B. & Hill, R. M. Evidence for a physiological explanation of the waterfall phenomenon and figural after-effects. *Nature* **200**, 1345–1347 (1963).

- Schieling, S. & Spillmann, L. Flicker adaptation in the peripheral retina. *Vision Res.* **27**, 277–284 (1987).
- Albrecht, D. G., Farrar, S. B. & Hamilton, D. B. Spatial contrast adaptation characteristics of neurones recorded in the cat's visual cortex. *J. Physiol.* **347**, 713–739 (1984).
- Ho, W. A. & Berkley, M. A. Evoked potential estimates of the time course of adaptation and recovery to counterphase gratings. *Vision Res.* **28**, 1287–1296 (1988).
- Giaschi, D., Douglas, R., Marlin, S. & Cynader, M. The time course of direction-selective adaptation in simple and complex cells in cat striate cortex. *J. Neurophysiol.* **70**, 2024–2034 (1993).
- Barlow, H. B. & Bridley, G. S. Inter-ocular transfer of movement aftereffects during pressure binding of the stimulated eye. *Nature* **200**, 1347 (1963).
- Meister, M., Lagnado, L. & Baylor, D. A. Concerted signaling by retinal ganglion cells. *Science* **270**, 1207–1210 (1995).
- Brainard, D. Calibration of a computer controlled color monitor. *Color Res. Appl.* **14**, 23–34 (1989).
- Dawis, S. M. Polynomial expressions of pigment nomograms. *Vision Res.* **21**, 1427–1430 (1981).
- Nuboer, J. F. W. & Moed, P. J. Increment-threshold spectral sensitivity in the rabbit. *J. Comp. Physiol.* **151**, 353–358 (1983).

Acknowledgements. We thank T. Jordan for help with rabbit experiments, the members of our group and B. Chang for discussion, and D. Baylor, J. Dowling, C. Reid, F. Rieke and J. Victor for comments on the manuscript. Supported by a grant from the Office of Naval Research and a Markey Scholarship to M.M., and a Research Assistantship from the Harvard–MIT Division of Health Sciences and Technology to S.S.

Correspondence and requests for materials should be addressed to M.M. (e-mail: meister@biosun.harvard.edu).

An endothelial receptor for oxidized low-density lipoprotein

Tatsuya Sawamura*, Noriaki Kume†, Takuma Aoyama*, Hideaki Moriawaki†, Hajime Hoshikawa*, Yuichi Aiba†, Takeshi Tanaka‡, Soichi Miwa*, Yoshimoto Katsura‡, Toru Kita† & Tomoh Masaki*

* Department of Pharmacology and † Department of Geriatric Medicine, Faculty of Medicine, Kyoto University, Kyoto 606, Japan

‡ Department of Immunology, Chest Disease Research Institute, Kyoto University, Kyoto 606, Japan

§ Japan Red Cross Saitama Branch, Yono, Saitama 338, Japan

Endothelial dysfunction or activation elicited by oxidatively modified low-density lipoprotein (Ox-LDL) has been implicated in the pathogenesis of atherosclerosis^{1–4}, characterized by intimal thickening and lipid deposition in the arteries. Ox-LDL and its lipid constituents impair endothelial production of nitric oxide, and induce the endothelial expression of leukocyte adhesion molecules and smooth-muscle growth factors, which may be involved in atherogenesis^{5–7}. Vascular endothelial cells in culture^{8,9} and *in vivo*^{10,11} internalize and degrade Ox-LDL through a putative receptor-mediated pathway that does not involve macrophage scavenger receptors^{12–15}. Here we report the molecular cloning, using expression cloning strategy, of an Ox-LDL receptor from vascular endothelial cells. The cloned receptor is a membrane protein that belongs structurally to the C-type lectin family, and is expressed *in vivo* in vascular endothelium and vascular-rich organs.

A cDNA library of cultured bovine aortic endothelial cells (BAEC), which bind, internalize and degrade Ox-LDL, was used for expression cloning. COS-7 cells transfected with a single clone, pBLOX-1, exhibited prominent uptake of DiI-labelled Ox-LDL. To characterize the protein encoded by pBLOX-1, designated lectin-like Ox-LDL receptor (LOX-1), as a receptor for Ox-LDL, we have transfected pBLOX-1 into CHO-K1 cells and established a cell line stably expressing bovine LOX-1 (BLOX-1-CHO). An antiserum was also developed using bacterially expressed extracellular domain of LOX-1 (61–270) as an antigen. Immunoblotting of total cell lysate of BLOX-1-CHO, using anti-LOX-1 antiserum, showed a single band with approximate relative molecular mass 50,000 (M_r 50K). This particular band was not detected in cell lysate from untransfected CHO-K1 cells (Fig. 1a). A band with 50K on the same position of the

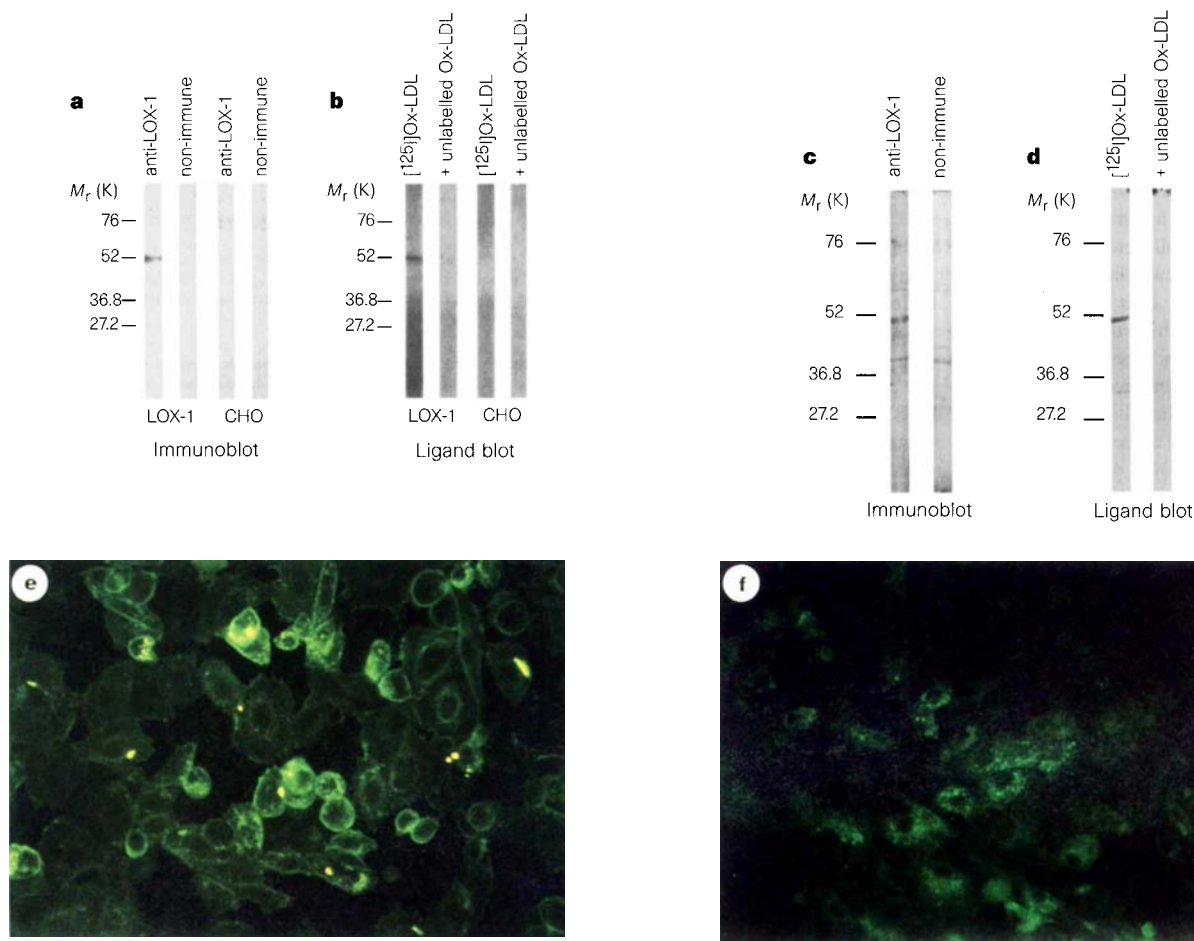


Figure 1 LOX-1 is a protein of M_r 50K that is expressed on the cell surface of bovine aortic endothelial cells. **a**, Immunoblotting of BLOX-1-CHO and untransfected CHO-K1 with anti-LOX-1 antiserum. A band at 50K was detected in BLOX-1-CHO, but not in untransfected CHO-K1. **b**, Ligand blotting of BLOX-1-CHO and untransfected CHO-K1 cells using [125 I]Ox-LDL as a ligand. A similar band of ~50K was detected in BLOX-1-CHO but not in untransfected CHO-K1. Addition of a 100-fold excess of unlabelled Ox-LDL abolished the binding of [125 I]Ox-LDL. **c**, Immunoblotting of BAEC by anti-LOX-1 antiserum. A band of 50K was detected in BAEC. A monoclonal antibody for bovine LOX-1 also detected the 50K band (data not shown). **d**, Ligand blotting of BAEC using [125 I]Ox-LDL as a ligand. A 50K

band was detected as described in immunoblotting of BAEC, as well as both immunoblotting and ligand blotting of BLOX-1-CHO. **e**, **f**, Indirect immunofluorescence of non-permeabilized BLOX-1-CHO (**e**) and BAEC (**f**) with anti-bovine LOX-1 antiserum. Cells were fixed with 3% paraformaldehyde, and subsequently incubated with anti-LOX-1 antibody and fluorescein isothiocyanate-conjugated anti-mouse IgG. Immunofluorescence microscopy shows that LOX-1 is expressed on the cell surface. Signals from the intracellular space of BLOX-1-CHO are leakage of the fluorescence of DiI-Ox-LDL incubated before cell fixation. Staining with preimmune serum did not show any significant signals (data not shown).

same membrane was detected by ligand blotting, using [125 I]-labelled Ox-LDL as a ligand (Fig. 1b). Both immunoblot and ligand blot of BAEC proteins similarly stained a single band on the same position of the same membrane, suggesting that this 50K protein binds Ox-LDL in BAEC (Fig. 1c, d). Indirect immunofluorescence of non-permeabilized BLOX-1-CHO cells showed that anti-LOX-1 antiserum bound to the cell surface of BLOX-1-CHO (Fig. 1e) and BAEC (Fig. 1f), but not to untransfected CHO-K1 cells (data not shown), demonstrating that LOX-1 is expressed on the cell surface.

To further confirm the function of LOX-1 as a receptor for Ox-LDL, binding and proteolytic degradation of [125 I]-labelled Ox-LDL, as well as internalization of DiI-labelled Ox-LDL, were examined. Binding of [125 I]-labelled Ox-LDL to both BLOX-1-CHO and BAEC were effectively blocked by 100-fold excess amounts of unlabelled Ox-LDL. The amounts of [125 I]-labelled Ox-LDL binding to BLOX-1-CHO cells were no less than those to CHO cells expressing bovine type II macrophage scavenger receptor (281 and 110 ng mg $^{-1}$ cellular protein, respectively)¹³. Two independent monoclonal antibodies directed to bovine LOX-1 significantly inhibited [125 I]-labelled Ox-LDL binding to both BLOX-1-

CHO and BAEC (Fig. 2). Fluorescence microscopy of BLOX-1-CHO incubated with DiI-labelled Ox-LDL showed that BLOX-1-CHO cells, but not untransfected CHO-K1 cells, internalized significant amounts of DiI-labelled Ox-LDL (Fig. 3a), which was also blocked by excess amounts of unlabelled Ox-LDL but not by unlabelled native LDL (data not shown). [125 I]-labelled Ox-LDL internalized into BLOX-1-CHO can be proteolytically degraded, as previously reported in cultured endothelial cells⁹. This degradation was effectively suppressed by excess amounts of unlabelled Ox-LDL but not by unlabelled native LDL (Fig. 3c). Polyinosinic acid, which inhibits endothelial binding of Ox-LDL¹¹, also significantly blocked binding of [125 I]-labelled Ox-LDL in BLOX-CHO. Thus LOX-1 seems to be expressed on the cell surface on BAEC, and can support binding, internalization and proteolytic degradation of Ox-LDL.

We screened a cDNA library constructed from human lung, and subsequently cloned cDNA encoding the human homologue of LOX-1. A cell line stably expressing human LOX-1 (HLOX-1-CHO) also exhibited the activity of the uptake of DiI-labelled Ox-LDL (Fig. 3b), which was displaced by unlabelled Ox-LDL but not by native LDL (data not shown). Degradation of [125 I]-labelled Ox-LDL by HLOX-1-CHO was also confirmed, and was inhibited by

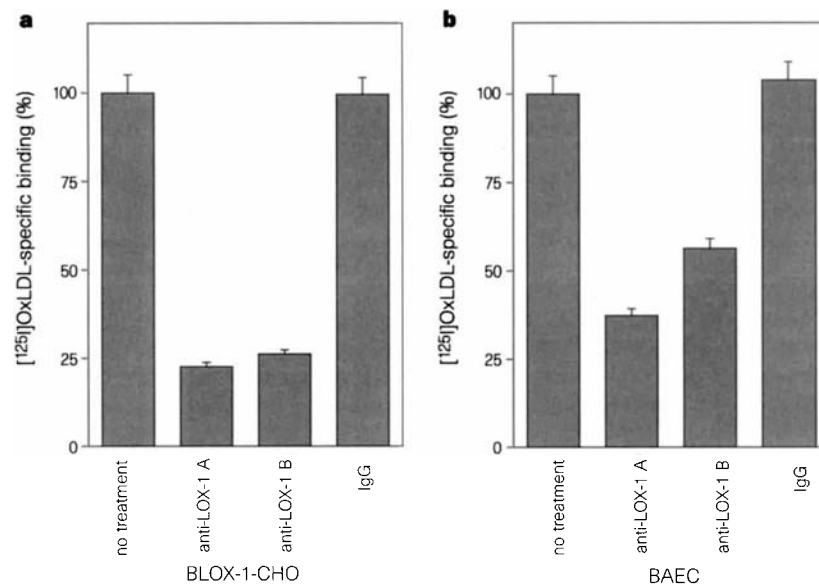


Figure 2 Binding of [¹²⁵I]Ox-LDL to BLOX-1-CHO (**a**) and BAEC (**b**). Cells were incubated with 5 $\mu\text{g ml}^{-1}$ of [¹²⁵I]Ox-LDL on ice for 1 h, washed 3 times, and cell-bound radioactivities were measured by a gamma counter. Specific binding of Ox-LDL was calculated as displaceable binding with 100-fold excess of (500 $\mu\text{g ml}^{-1}$) unlabelled Ox-LDL. Two independent anti-BLOX-1 monoclonal antibodies (30 $\mu\text{g ml}^{-1}$) or a comparable amount of non-immune mouse IgG

was also included in incubation media. Addition of monoclonal antibodies directed to LOX-1, but not control non-immune IgG, significantly reduced the specific binding of [¹²⁵I]Ox-LDL to BAEC. The absolute values of specific binding of [¹²⁵I]Ox-LDL to BLOX-1-CHO and BAEC were 281 ± 7.8 per ng mg protein and 20 ± 1.2 ng per mg protein (mean $\pm$ s.e.m.), respectively.

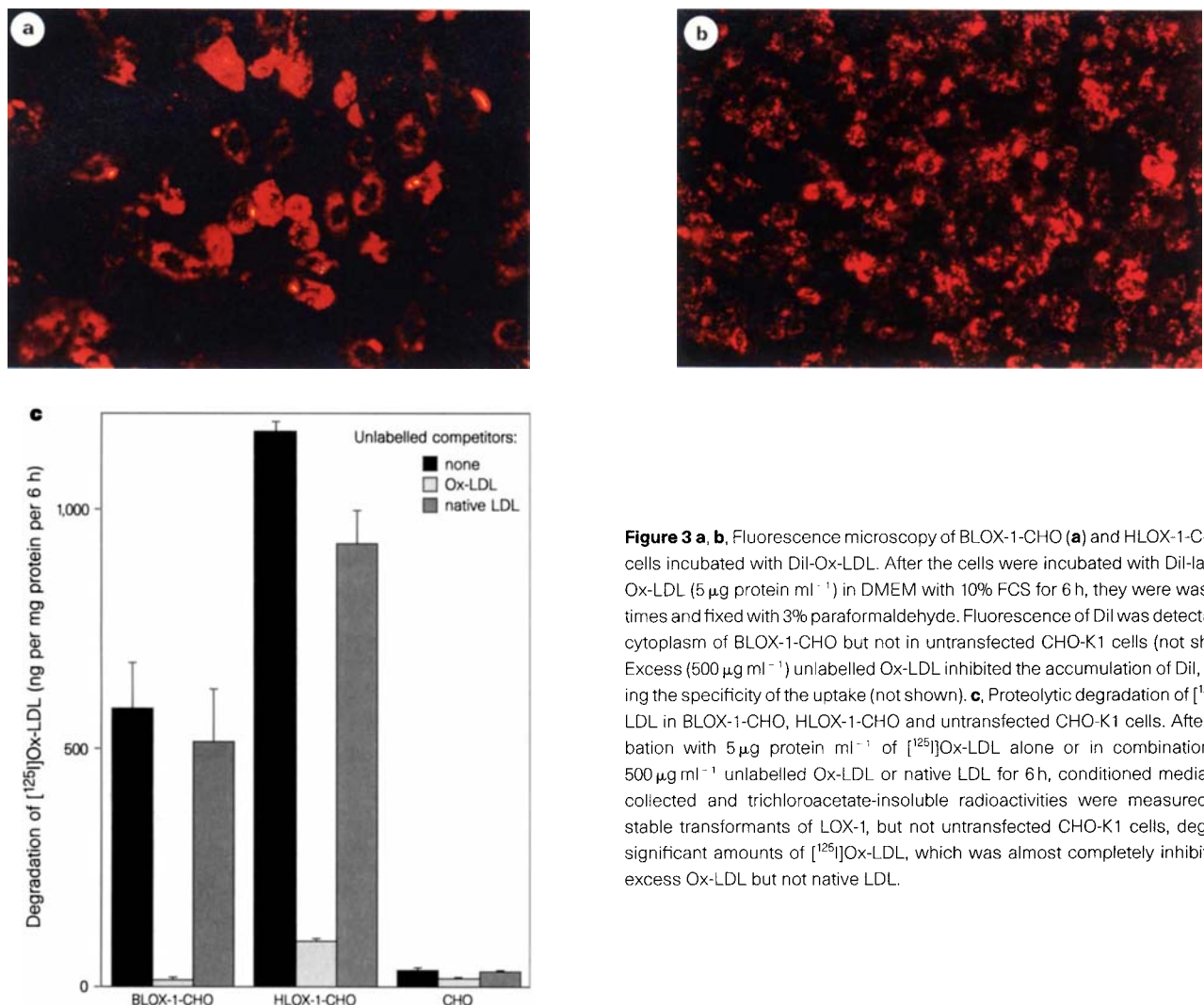


Figure 3 a, b, Fluorescence microscopy of BLOX-1-CHO (**a**) and HLOX-1-CHO (**b**) cells incubated with Dil-Ox-LDL. After the cells were incubated with Dil-labelled Ox-LDL (5 $\mu\text{g protein ml}^{-1}$) in DMEM with 10% FCS for 6 h, they were washed 3 times and fixed with 3% paraformaldehyde. Fluorescence of Dil was detectable in cytoplasm of BLOX-1-CHO but not in untransfected CHO-K1 cells (not shown). Excess (500 $\mu\text{g ml}^{-1}$) unlabelled Ox-LDL inhibited the accumulation of Dil, showing the specificity of the uptake (not shown). **c**, Proteolytic degradation of [¹²⁵I]Ox-LDL in BLOX-1-CHO, HLOX-1-CHO and untransfected CHO-K1 cells. After incubation with 5 $\mu\text{g protein ml}^{-1}$ of [¹²⁵I]Ox-LDL alone or in combination with 500 $\mu\text{g ml}^{-1}$ unlabelled Ox-LDL or native LDL for 6 h, conditioned media were collected and trichloroacetate-insoluble radioactivities were measured. The stable transformants of LOX-1, but not untransfected CHO-K1 cells, degraded significant amounts of [¹²⁵I]Ox-LDL, which was almost completely inhibited by excess Ox-LDL but not native LDL.

unlabelled Ox-LDL but not by native LDL (Fig. 3c). These similar results with the human homologue of LOX-1 demonstrate that LOX-1 works as a receptor for Ox-LDL in both species.

The cDNA for bovine LOX-1 cloned from BAEC contained an open reading frame of 810 base pairs that encodes a protein of 270 amino-acid residues with a calculated M_r of 30,872. The expressed protein in CHO-K1 cells and BAEC had an M_r of 50K (Fig. 1a, c), which may result from glycosylation of four potential N-linked glycosylation sites located on C-terminal domain. In the amino-terminal cytoplasmic domain, several potential phosphorylation sites were found (Thr 21 and Ser 28 for protein kinase C; Thr 2 for casein kinase II), suggesting that phosphorylation of these sites may transmit biological signals or regulate the function of LOX-1.

Human LOX-1 has 273 amino acids, and an M_r of 30,939. The structure of human LOX-1 was homologous to bovine LOX-1, with 72% identity in the amino-acid sequences. Homology was further conserved in the lectin domain (see below); 81% was identical in the amino-acid sequences (Fig. 4a).

LOX-1 has the structure conserved in the C-type lectin family¹⁶. It has a type II membrane protein structure with a short N-terminal hydrophilic and a long carboxy-terminal hydrophilic domain, separated by a hydrophobic domain of 26 amino acids. Six repeats of cysteines in the homologous lectin domain were also found. LOX-1 has the highest homology with NKR-P1 family proteins expressed on the surface of natural killer (NK) cells¹⁷, which are involved in target-cell recognition and NK-cell activation (Fig. 4b).

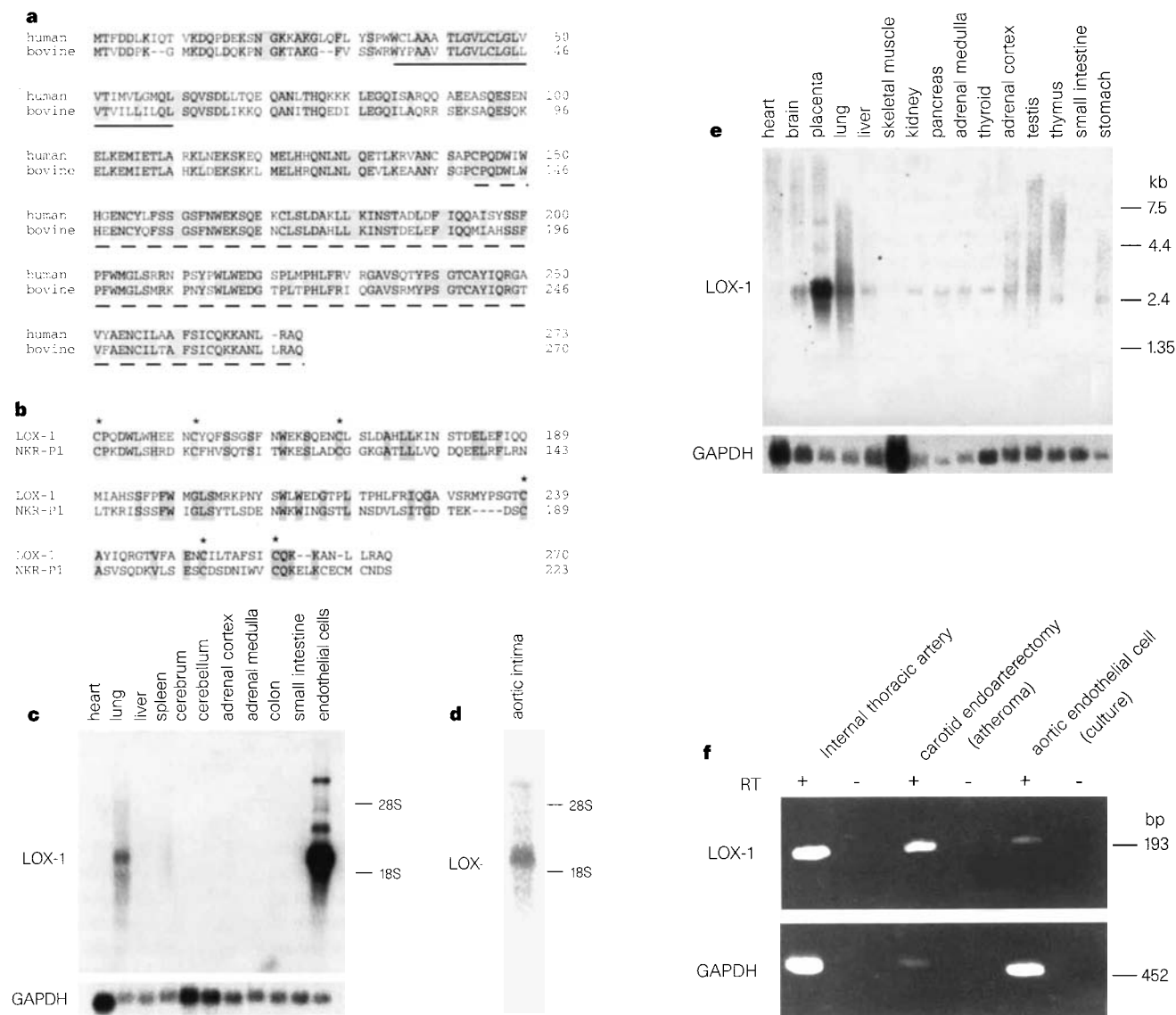


Figure 4 **a**, Comparison of amino-acid sequences between human and bovine LOX-1. Conserved amino-acid sequences in both species are shaded. The putative transmembrane domain is underlined, and the C-type lectin domain is indicated by dashed underline. **b**, Alignment of C-type lectin domains of LOX-1 and rat NKR-P1. The positions of six cysteine residues are conserved (asterisks). **c-e**, Northern blot analyses of bovine tissues and cultured aortic endothelial cell (**c**), bovine aortic intima (**d**), and human tissues (**e**). LOX-1 mRNA was detected as bands for 2.3 kb and 2.8 kb in bovine and human tissues, respectively. To control the amounts of mRNA loaded, blots were reprobed with cDNA encoding glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Note that the exposure

of bovine membrane (**c**) was shorter than that of human membrane (**e**) because of the high level of expression of LOX-1 mRNA in cultured bovine aortic endothelial cells. **f**, RT-PCR of human internal thoracic artery, a specimen from carotid endoarterectomy, and cultured human aortic endothelial cells. LOX-1 cDNA fragments with 193 base pairs were amplified by primer pair for human LOX-1 cDNA. To control the amounts of mRNA 452 base-pair fragments of GAPDH cDNA were amplified using Amplimer set (Clontech) from the same RNA preparations. These cDNA fragments were undetectable when reverse transcription was omitted.

LOX-1, like NKR-P1 might also transmit some biological signals and thereby activate endothelial cells after ligation to Ox-LDL. LOX-1 does not have any homology with known receptors for Ox-LDL which were mainly found in macrophages^{13,14,18–21}. Therefore, molecular mechanisms of binding of this C-type lectin-like molecule to Ox-LDL may be distinct from those previously found for Ox-LDL receptors, and remain to be clarified. The C-type lectin family includes 'selectins' that are involved in cell-cell recognition, including endothelial-leukocyte adhesion. Macrophage class A scavenger receptors were involved in divalent cation-independent adhesion of macrophages to substrates²². LOX-1 may also, therefore, have additional (patho)physiological functions, including cell-cell and cell-matrix adhesion.

To define whether LOX-1 is expressed *in vivo*, as well as in cultured endothelial cells, northern blot analyses and the polymerase chain reaction coupled with reverse transcription (RT-PCR) were performed. Northern blot analyses of various tissues revealed that aortic intima and vascular-rich organs, such as placenta, lungs, brain and liver, express LOX-1 mRNA *in vivo* in physiological conditions (Fig. 4c–e). RT-PCR showed that human LOX-1 is expressed in normal thoracic and carotid arterial cells including atheromatous lesions, as well as cultured aortic endothelial cells (Fig. 4f). The 3'-non-coding region of bovine LOX-1 cDNA contains seven mRNA unstabilizing signals, AUUUA²³, indicating that mRNA for LOX-1 may have a short intracellular half-life. This suggests that expression of LOX-1 might be dynamically regulated by certain biological stimuli that are relevant to atherogenesis. Certain inflammatory cytokines, Ox-LDL and its atherogenic lipid constituent can upregulate the expression of LOX-1 (data not shown). Although the potential roles of LOX-1 in atherogenesis are not yet fully understood, Ox-LDL uptake through this receptor expressed on the surface of vascular endothelium may be involved in endothelial activation or dysfunction in atherogenesis. □

Methods

Preparation of Ox-LDL. Human plasma LDLs (relative density 1.019–1.063) were isolated by sequential ultracentrifugation, and oxidative modification of LDL was performed with Cu²⁺ *in vitro*. Oxidation was monitored by measuring the amount of thiobarbituric acid-reactive substances (around 10 nmol malondialdehyde equivalent per mg protein in Ox-LDL)⁹. Agarose gel electrophoresis showed increased electrophoretic mobility and minimal aggregation of Ox-LDL particles. LDL was labelled with 1,1'-dioctadecyl-3,3',3'-tetramethylindocarbocyanine perchlorate (DiI)²⁴ (Molecular Probes).

cDNA cloning of bovine LOX-1. A cDNA library of bovine endothelial cells was constructed in a mammalian expression vector pME18S²⁵. The plasmid library consisted of 7 × 10⁵ independent clones. The library was once amplified and used for transfection.

For COS-7 transfection and cell sorting, 10 µg of plasmids were transfected into COS-7 cells by using Lipofectamine (Gibco). COS-7 cells were incubated 48 h after transfection with 3 µg protein per ml of DiI-labelled Ox-LDL for 12 h. DiI-positive COS-7 cells were sorted by a FACS Vantage cell sorter (Becton Dickinson). Episomal plasmids were collected from the sorted cells²⁶ and electroporated into DH10B. Plasmid DNAs were purified after amplification and subjected to another round of transfection and sorting. After three rounds of selection, plasmids from single clones were prepared and individually transfected into COS-7 cells.

Stable transformant of LOX-1. pBLOX-1 (1 µg) and pSV2bssr (10 ng) (Funakoshi, Japan) were co-transfected into CHO-K1 cells by the calcium-phosphate transfection method. Stable transformants were selected under the existence of 10 µg ml⁻¹ of blasticidin S (Funakoshi).

Antibodies against LOX-1. A cDNA fragment covering the extracellular domain of bovine LOX-1 (amino acids 61–270) was amplified by PCR with a pair of primers tagged with a *Bam*HI restriction site. The amplified fragment was digested with *Bam*HI and subcloned into the *Bam*HI site of pQE10 vector (QIAGEN). Protein synthesis and purification of the extracellular domain of LOX-1 were performed using the QIAexpress system (QIAGEN). The protein was used as the antigen for immunizing mice. Hybridomas producing monoclonal

antibodies were made by standard procedures and screened by enzyme-linked immunosorbent assay. Two independent clones that showed significant inhibition in the binding of [¹²⁵I]-Ox-LDL to BLOX-1-CHO were selected.

Immunoblot and ligand blot analyses. Cells were solubilized directly in the sample buffer for SDS-PAGE without reducing reagents. The extracts were separated by SDS-PAGE in non-reducing conditions and blotted onto nylon membranes. After blocking with Block Ace (Snow Brand, Japan), the membranes were incubated with 10 µg ml⁻¹ [¹²⁵I]-Ox-LDL with or without 1 mg ml⁻¹ unlabelled Ox-LDL in 50 mM Tris-HCl, pH 8.0, 2 mM CaCl₂, 90 mM NaCl, 25% (v/v) Block Ace at room temperature. The membranes were washed in the same buffer without ligands and autoradiographed with BAS2000 system (Fuji). Immunostaining of the membranes with an anti-bovine LOX-1 antibody was performed using biotinylated second antibody and peroxidase-conjugated avidin-biotin complex (Vector) and Immunostain kit (Konica).

cDNA cloning of human LOX-1. A cDNA library of human lung primed with random primer and oligo(dT) primer was screened by full-length bovine LOX-1 cDNA. A single positive clone (hLOX-1) was used which covers human LOX-1 (465–1192). A 5'-upstream region of human LOX-1 cDNA was obtained from poly(A)⁺ RNA from human placenta by 5'-RACE using specific primer corresponding to the antisense strand of human LOX-1 (555–581) (ref. 27). The cDNA sequence was confirmed by three independent clones. Sequence homology was analysed by Geneworks (Intelligenetics).

Northern blot analyses. Poly(A)⁺ RNA (5 µg per lane), prepared from bovine tissues, was separated by a formaldehyde/1.1% agarose gel electrophoresis, and transferred to a nylon membrane (Gene Screen Plus, DuPont). These membranes and multiple tissue northern blot of human tissues (Clontech) were hybridized with cDNA of bovine and human LOX-1 labelled with [α -³²P]dCTP by random priming method, respectively. Hybridization was performed at 42 °C in 2 × SSC, 50% formamide, 1% SDS, 0.1 mg ml⁻¹ salmon sperm DNA, 10% dextran sulphate. The membrane was washed in 0.2 × SSC/0.1% SDS at 65 °C, and autoradiographed with BAS2000 system (Fuji).

RT-PCR. Total RNA (500 ng) extracted from fragments of human arteries was reverse transcribed with random hexamer using Super Script (Gibco). As much as 5% of the reverse-transcribed materials were amplified with LA-Taq DNA polymerase (Takara) using a primer pair specific to human LOX-1 (sense primer, 5'-TTACTCTCCATGGTGGTGCC-3', antisense primer, 5'-AGCTTCTCTGCTTGTGGCC-3'). For PCR, 35 cycles were used at 94 °C for 40 s, 55 °C for 1 min, and 72 °C for 1 min.

Received 16 September 1996; accepted 6 January 1997.

1. Brown, M. S. & Goldstein, J. L. *Annu. Rev. Biochem.* **52**, 223–261 (1983).
2. Steinberg, D., Parthasarathy, S., Carew, T. E., Khoo, J. C. & Witztum, J. L. *N. Engl. J. Med.* **320**, 915–924 (1989).
3. Kita, T. *et al. Proc. Natl. Acad. Sci. USA* **84**, 5928–5931 (1987).
4. Ross, R. *Nature* **368**, 801–809 (1993).
5. Gimbrone, M. A. J., Cybulsky, M. I., Kume, N., Collins, T. & Resnick, N. *Ann. NY Acad. Sci.* **748**, 122–132 (1995).
6. Kume, N. & Gimbrone, M. A. J. *J. Clin. Invest.* **93**, 907–911 (1994).
7. Kugiyama, K., Kerns, S. A., Morrisett, J. D., Roberts, R. & Henry, P. D. *Nature* **344**, 160–162 (1990).
8. Stein, O. & Stein, Y. *Biochim. Biophys. Acta* **620**, 631–635 (1980).
9. Kume, N., Arai, H., Kawai, C. & Kita, T. *Biochim. Biophys. Acta* **1091**, 63–67 (1991).
10. Pitas, R. E., Boyles, J., Mahley, R. W. & Bissel, D. M. *J. Cell. Biol.* **100**, 103–117 (1985).
11. van Berkel, T. J. C., De Rijke, Y. B. & Kruijff, J. K. *J. Biol. Chem.* **266**, 2282–2289 (1991).
12. Krieger, M. *et al. J. Biol. Chem.* **268**, 4569–4572 (1993).
13. Kodama, T. *et al. Nature* **343**, 531–535 (1990).
14. Rohrer, L., Freeman, M., Kodama, T., Penman, M. & Krieger, M. *Nature* **343**, 570–572 (1990).
15. Bickel, P. E. & Freeman, M. W. *J. Clin. Invest.* **90**, 1450–1457 (1992).
16. Drickamer, K. J. *J. Biol. Chem.* **263**, 9557–9560 (1988).
17. Giorda, R. *et al. Science* **249**, 1298–1300 (1990).
18. Endemann, G. *et al. J. Biol. Chem.* **268**, 11811–11816 (1993).
19. Acton, S. L., Scherer, P. E., Lodish, H. F. & Krieger, M. *J. Biol. Chem.* **269**, 21003–21009 (1994).
20. Standon, L. W., White, R. T., Bryant, C. M., Protter, A. A. & Endemann, G. *J. Biol. Chem.* **267**, 22446–22451 (1992).
21. Ramprasad, M. P. *et al. Proc. Natl. Acad. Sci. USA* **92**, 9580–9584 (1995).
22. Fraser, I., Hughes, D. & Gordon, S. *Nature* **364**, 343–346 (1993).
23. Show, G. & Kamen, R. *Cell* **46**, 659–667 (1986).
24. Stephan, Z. F. & Yurachek, E. C. *J. Lipid Res.* **34**, 325–330 (1993).
25. Ikura, T. *et al. Biochem. Biophys. Res. Commun.* **203**, 1417–1422 (1994).
26. Hirt, B. *J. Mol. Biol.* **26**, 365–369 (1967).
27. Frohman, M. A., Dush, M. K. & Martin, G. R. *Proc. Natl. Acad. Sci. USA* **85**, 8998–9002 (1988).

Acknowledgements. We thank T. Kodama for the stable cell line expressing bovine type II scavenger receptor. This work was supported in part by grants-in-aid for scientific research from the Ministry of Education, Science and Culture of Japan.

Correspondence and requests for materials should be addressed to T.M. (e-mail: masaki@mfour.med.kyoto-u.ac.jp). The sequences for bovine and human LOX-1 have been deposited in DDBJ, EMBL and GenBank nucleotide sequence databases, accession nos D89049 and D89050, respectively.

Spontaneous calcification of arteries and cartilage in mice lacking matrix GLA protein

Guangbin Luo*, Patricia Ducey*, Marc D. McKee†, Gerald J. Pinero‡, Evelyn Loyer§, Richard R. Behringer* & Gérard Karsenty*

Departments of* Molecular Genetics, and § Radiology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, 77030, USA

† Department of Stomatology, Faculty of Dentistry, Université de Montréal, Montréal, Québec H3C 3J7, Canada

‡ Department of Basic Science, Dental Branch-University of Texas, Houston, Texas 77030, USA

Calcification of the extracellular matrix (ECM) can be physiological or pathological. Physiological calcification occurs in bone when the soft ECM is converted into a rigid material capable of sustaining mechanical force; pathological calcification can occur in arteries¹ and cartilage² and other soft tissues. No molecular determinant regulating ECM calcification has yet been identified. A candidate molecule is matrix GLA protein (Mgp), a mineral-binding ECM protein³ synthesized by vascular smooth-muscle cells and chondrocytes, two cell types that produce an uncalcified ECM. Mice that lack Mgp develop to term but die within two months as a result of arterial calcification which leads to blood-vessel rupture. Chondrocytes that elaborate a typical cartilage matrix can be seen in the affected arteries. Mgp-deficient mice additionally exhibit inappropriate calcification of various cartilages, including the growth plate, which eventually leads to short stature, osteopenia and fractures. These results indicate that ECM calcification must be actively inhibited in soft tissues. To our knowledge, Mgp is the first inhibitor of calcification of arteries and cartilage to be characterized *in vivo*.

Mgp belongs to the family of mineral-binding GLA proteins that includes coagulation factors VII and IX, the ant clotting factors protein S and protein C⁴, and osteocalcin, an inhibitor of bone

formation⁵. GLA proteins bind mineral and mineral ions through γ -carboxylated glutamic acid residues⁶. To investigate Mgp function *in vivo*, we generated mice with a disrupted *Mgp* allele by gene targeting in embryonic stem (ES) cells (Fig. 1a–c). Heterozygous mice, which were viable and fertile, were intercrossed to generate homozygous Mgp-deficient mice (*mgp*^{ml}/*mgp*^{ml}). These mice were born alive at the expected mendelian ratio. No Mgp expression was detected in the *mgp*^{ml}/*mgp*^{ml} mice (Fig. 1d), indicating that *mgp*^{ml} is a null allele. These Mgp-deficient mice displayed no phenotypic changes within the first two weeks of life. Thereafter, they could easily be distinguished from their littermates: they had a noticeably faster heart beat, became shorter (Fig. 1e) and died within two months.

Autopsies of the mutant animals showed that the cause of death was haemorrhage due to rupture of the thoracic or abdominal aorta. Alizarin red staining for mineralized tissues revealed extensive calcification along the aorta and its branches. The calcification originated at multiple sites and affected all elastic and muscular arteries, but not the arterioles, capillaries or veins (Fig. 2a). Histology of arteries in *mgp*^{ml}/*mgp*^{ml} mice of up to 1 week old showed no differences compared with wild-type mice (data not shown). In contrast, in 2-week-old (data not shown) and 3-week-old *mgp*^{ml}/*mgp*^{ml} mice (Fig. 2b), von Kossa staining for mineral revealed prominent calcification of the elastic lamellae in the media of the aortic wall and disruption of the normal tissue architecture of the blood vessel. There was calcification in the coronary arteries (Fig. 2c) and aortic valves (data not shown) but not in the myocardium, and no fatty streak or atherosclerotic plaques¹ in affected arteries. Transmission electron microscopy (TEM) revealed electron-dense mineral in elastic fibres, which had also infiltrated nearby collagen fibrils (Fig. 2d). Electron diffraction identified the inorganic material as apatite, the mineral of bone and of calcified atherosclerotic lesions⁷ (Fig. 2d, right inset).

Analysis of affected arteries in the oldest animals showed the existence of cells with chondrocytic features, including hypertrophic chondrocytes surrounded by a typical metachromatic cartilage matrix⁸ (Fig. 2f). TEM confirmed the chondrocyte-like morphology of these cells and the cartilaginous nature of the ECM, indicated by the presence of type II collagen fibrils and proteoglycans (Fig. 2e), two hallmarks of cartilage ECM⁸. Matrix vesicles,

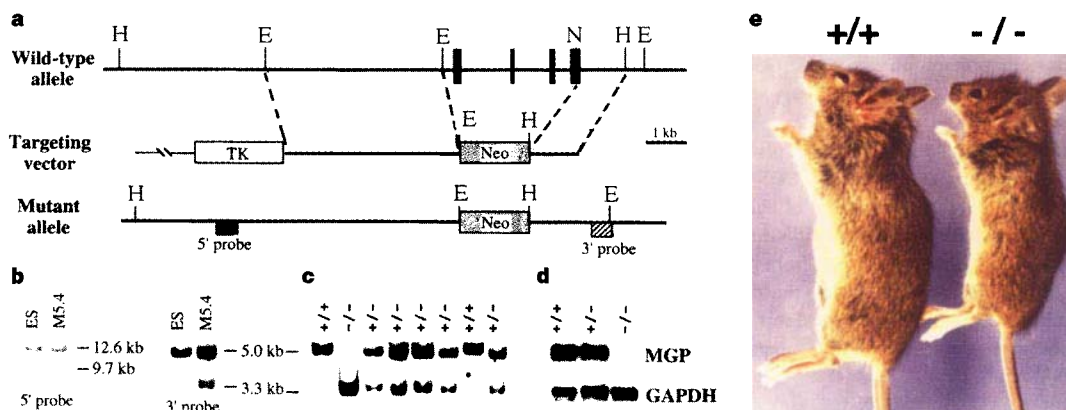


Figure 1 Gene targeting at the *Mgp* locus. **a**, The expected gene replacement at the mouse *Mgp* locus. Correct targeting should result in the replacement of a 3.4-kb region, containing the proximal promoter, exons 1, 2, 3 and most of 4 of the *Mgp* gene³⁰, by a marker gene for positive selection (PGKNeo)³⁴. The MC1-TK (thymidine kinase) expression cassette²⁸ was used for negative selection. The 5' and 3' external probes used for Southern blot analysis are shown as patterned boxes. E, *EcoRI*; H, *HindIII*; N, *NcoI*. **b**, Southern blot analysis of targeted ES cell DNA. The presence of a 9.7-kb *HindIII* fragment (5' probe) and a 3.3-kb *EcoRI* fragment (3' probe) indicates proper targeting at the *Mgp* locus. **c**, Southern blot analysis of

DNA from offspring derived from heterozygous matings. Genomic DNA isolated from the tails of offspring from one litter was digested with *EcoRI* and analysed using the 3' probe. The presence of a single 3.3-kb fragment is indicative of a homozygous mutant (*mgp*^{ml}/*mgp*^{ml}). **d**, Northern blot of RNA from offspring derived from heterozygous matings. 15 μ g total RNA collected from lung was probed as indicated. The absence of any hybridized band in the lane genotyped as *mgp*^{ml}/*mgp*^{ml} (–/–) confirmed that *mgp*^{ml} is a null allele. **e**, Morphological analysis of 5-week-old wild-type (+/+) and *mgp*^{ml}/*mgp*^{ml} (–/–) littermates. Note the shorter stature of the mutant mouse.

thought to be an initial site of calcification in the skeleton and in atherosclerotic lesions^{9,10}, were found in the vicinity of these chondrocytes (inset in Fig. 2e).

The *mgp*^{m1}/*mgp*^{m1} mice had a normal size at birth and for the first 2–3 weeks of life. Thereafter, they grew more slowly than wild-type littermates, and at the end of their lifespan were significantly shorter than wild-type mice (Fig. 1e). Von Kossa staining (Fig. 3a, b) and TEM (data not shown) were used to determine whether the short stature of the *mgp*^{m1}/*mgp*^{m1} mice was a result of inappropriate calcification of the growth-plate cartilage. Both techniques revealed that calcification normally restricted to the lower hypertrophic zone in wild-type animals⁸ extended into the zone of proliferating chondrocytes in the mutant animals. In physiological situations, the proper organization of chondrocytes in columns in the growth plate, and the control of cartilage matrix production provide the scaffolding necessary for bone matrix deposition by osteoblasts and for skeletal growth⁸. Instead, the abnormal calcification of the growth plate observed in the *mgp*^{m1}/*mgp*^{m1} mice led to a disorganization of chondrocyte columns (Fig. 3d), resulting in short stature (Fig. 1e), osteopenia (Fig. 3f) and fractures (data not shown). In addition, other cartilages containing proliferating chondrocytes,

such as the lower end of the trachea and the main bronchi, were calcified in the mutant mice, as indicated by alizarin-red staining of skeletal preparations (Fig. 2a). Von Kossa staining revealed the presence of mineral extending throughout the cartilage and surrounding tracheal chondrocytes (data not shown). The combined vascular and skeletal abnormalities in *mgp*^{m1}/*mgp*^{m1} mice mimic the phenotype of a rare human disease, Singleton–Merten syndrome¹¹, characterized by calcification of arteries without atherosclerosis, short stature, osteopenia of the extremities, and death in the first decade of life.

The two phenotypic abnormalities of *mgp*^{m1}/*mgp*^{m1} mice indicate that *Mgp* inhibits calcification of certain ECMs, probably through its mineral-ion-binding ability. These two phenotypes occur in the two tissues that express *Mgp* (Fig. 4). Starting at embryonic day 12.5 (E12.5), throughout development and after birth, *Mgp* is highly expressed in the aortic valves and in the media of arteries, but is not expressed in other non-vascular smooth muscle cells or in myocardium (data not shown, and Fig. 4). In cartilage it is strongly expressed in proliferating chondrocytes of the growth-plate cartilage and of the trachea and, to a lesser extent, in the late hypertrophic chondrocytes present in the growth plate. Although *Mgp* is

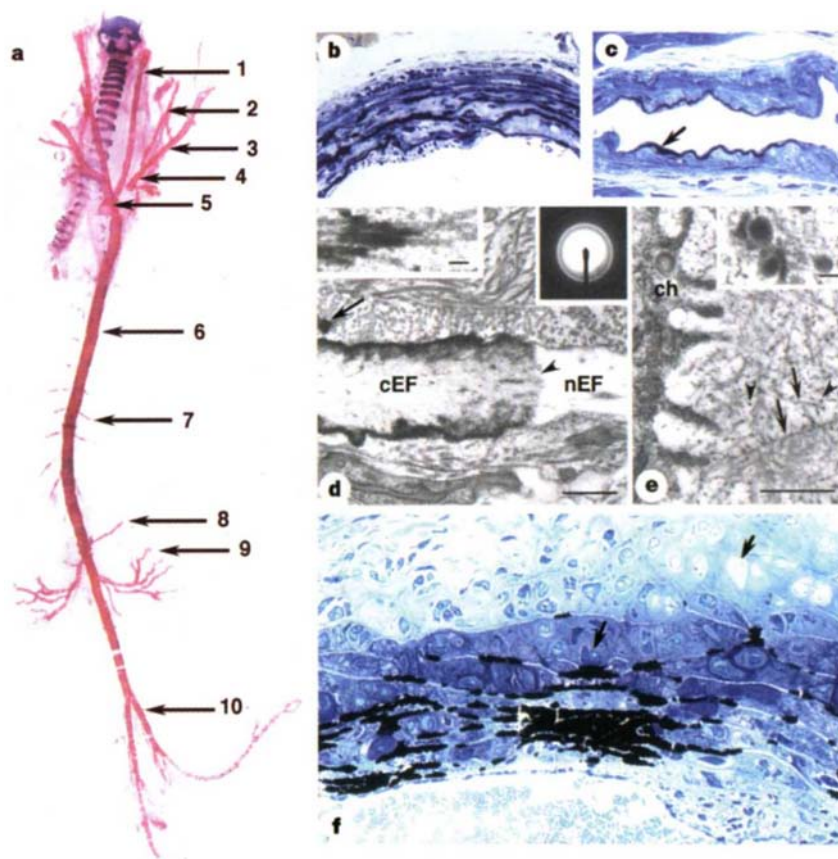


Figure 2 Morphological and histological analysis of blood vessels in *mgp*^{m1}/*mgp*^{m1} mice. **a**, Arteries of a 4-week-old *mgp*^{m1}/*mgp*^{m1} mouse dissected from a skeletal preparation stained with alizarin red for mineral. Calcified arteries resist the alkaline digestion of the soft tissues, whereas veins are completely digested. 1, Carotid artery; 2, cervical trunk; 3, axillary artery; 4, unnamed artery; 5, aortic arch; 6, aorta; 7, intercostal artery; 8, coeliac artery; 9, renal artery; 10, iliac artery. **b, c**, Von Kossa staining of arteries of 4-week-old *mgp*^{m1}/*mgp*^{m1} mice. Aorta of *mgp*^{m1}/*mgp*^{m1} is intensively stained black for mineral within elastic lamellae, and smooth-muscle cells have lost their spatial organization as discrete, contiguous layers throughout the blood vessel wall (**b**). Coronary artery from an *mgp*^{m1}/*mgp*^{m1} mouse showing calcification of the internal elastic lamina (arrow). **d, e**, Ultrastructural analysis (TEM) of calcified aortae of *mgp*^{m1}/*mgp*^{m1} mice.

d, Extensive calcification of elastic fibres (cEF) and of collagen fibrils (arrow, and left inset) occurs in the media of the aortic wall. Calcification advances along elastic fibres via a mineralization front (arrowhead). Selected-area electron diffraction of the mineral found in the elastic lamellae reveals diffraction reflections (002, 211, 112, 300) typical of apatite (right inset). nEF, normal elastic fibre. Scale bars, 0.5 and 0.1 μ m (inset). **e**, Chondrocytes (ch) surrounded by hyaline cartilage and type II collagen fibrils (arrows), proteoglycans (arrowheads) and matrix vesicles (inset in **e**) are present in the media of affected arteries. Scale bars, 1 μ m (**b**), 0.5 μ m (**c**), and 0.1 μ m (inset). **f**, Aorta of an *mgp*^{m1}/*mgp*^{m1} mouse displaying cartilaginous metaplasia, marked by the presence of chondrocytes (arrows) and metachromatic cartilage matrix.

expressed early during embryogenesis, we observed no defects in newborn or 1-week-old *mgp^{mi}/mgp^{mi}* mice, indicating that other gene product(s) might function to prevent calcification earlier during development.

The *mgp^{mi}/mgp^{mi}* phenotypes have important implications for our understanding of the consequences of physiological and pathological calcification in vertebrates. Calcification is seen in many arterial degenerative diseases, where it alters blood-vessel flexibility and promotes thrombosis and arterial rupture¹²⁻¹⁴. *mgp^{mi}/mgp^{mi}* mice have arterial calcification but no evidence of atherosclerotic plaques or other lesions, suggesting that genes like *Apo-E* may control the formation of atherosclerotic lesions¹⁵⁻¹⁸, whereas others like *Mgp* control the calcification process. This is consistent with the presumed genetic complexity of degenerative arterial diseases^{1,19,20}. The cellular and molecular events that cause skeletal growth to stop in vertebrates are not understood. A gene coding for the secreted protein parathyroid-hormone-related peptide, whose expression is controlled by *Indian Hedgehog* in the limbs, regulates differentiation of proliferative chondrocytes into hypertrophic chondrocytes and therefore the extent of longitudinal bone growth^{21,22}. The skeletal phenotype of the *mgp^{mi}/mgp^{mi}* mice indicates that another mechanism involving inhibition of calcification of cartilage ECM by *Mgp* may maintain the columnar organization of chondrocytes necessary for longitudinal bone growth. Finally, our demonstration that inhibition of calcification requires a

specific gene product in arteries and in cartilage suggests that calcification is a ubiquitous process and that other GLA proteins may prevent ECM calcification in organs that do not normally calcify and in which *Mgp* is not expressed. □

Methods

Gene targeting. Culture and transfection of ES cells have been described²³. Briefly, AB2.1 ES cells were transfected with 25 µg of linearized targeting vector per 10⁷ cells using a Biorad Gene Pulser and grown under double selection as described^{24,25}. Targeted ES cell clones were identified by Southern blot hybridization and were injected into C57BL/6J blastocysts²³. Male chimaeras were mated to C57BL/6J females, and heterozygous offspring were intercrossed to generate *mgp^{mi}/mgp^{mi}* mice.

Skeleton preparation. Skeletons were prepared as described²⁶. Four-week-old mice were killed, skinned, partly eviscerated, and fixed in 95% ethanol. Soft tissues were digested with 2% potassium hydroxide. Cartilaginous and calcified structures were stained with alcian blue 8GX and alizarin red S, respectively.

Histological analysis. Specimens for histological analysis were processed as described²⁷. Tissues were fixed with glutaraldehyde and paraformaldehyde, post-fixed with osmium tetroxide and embedded in epoxy (Epon) or acrylic (LR-White) resin. For light microscopy, semi-thin sections were left undecalcified and stained with 3% silver nitrate (von Kossa's reagent) followed by a 1% toluidine blue counterstaining. For electron microscopy analysis, thin sections were either left unstained for mineral analysis and selected-area electron

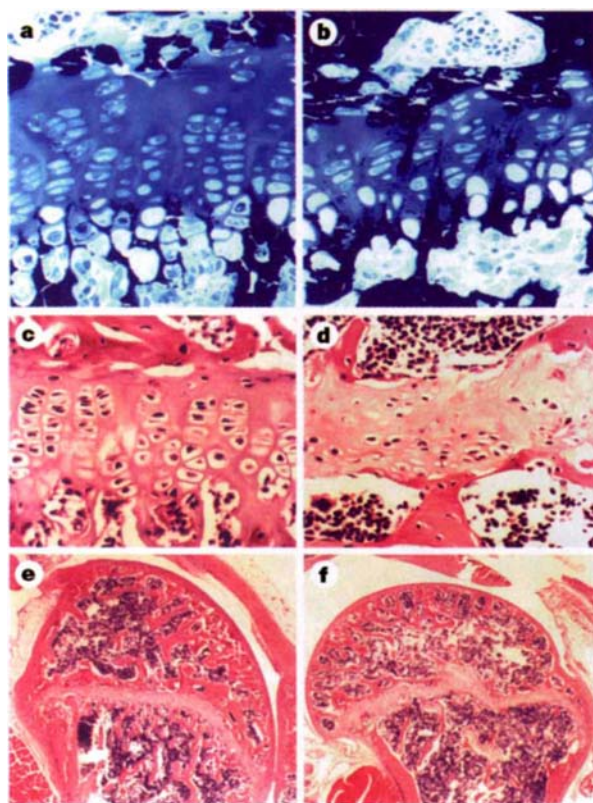


Figure 3 Histological analysis of cartilage of *mgp^{mi}/mgp^{mi}* mice. **a, b**, Von Kossa staining of growth-plate cartilage in 8-week-old wild-type (**a**) and *mgp^{mi}/mgp^{mi}* (**b**) mice. Cartilage in the proliferating chondrocyte zone of the *mgp^{mi}/mgp^{mi}* mouse is intensely stained by the von Kossa reaction for mineral (black areas), whereas no calcification is visible in the same area of the wild-type growth plate. **c-f**, Histological analysis of growth-plate cartilage and femoral head of 12-week-old wild-type (**c, e**) and *mgp^{mi}/mgp^{mi}* (**d, f**) mice. Proliferating chondrocytes are not organized in columns as in wild-type growth plate, and hypertrophic chondrocytes are absent in the mutant mice (**d**). Note the osteopenia in the mutant animal (**f**).

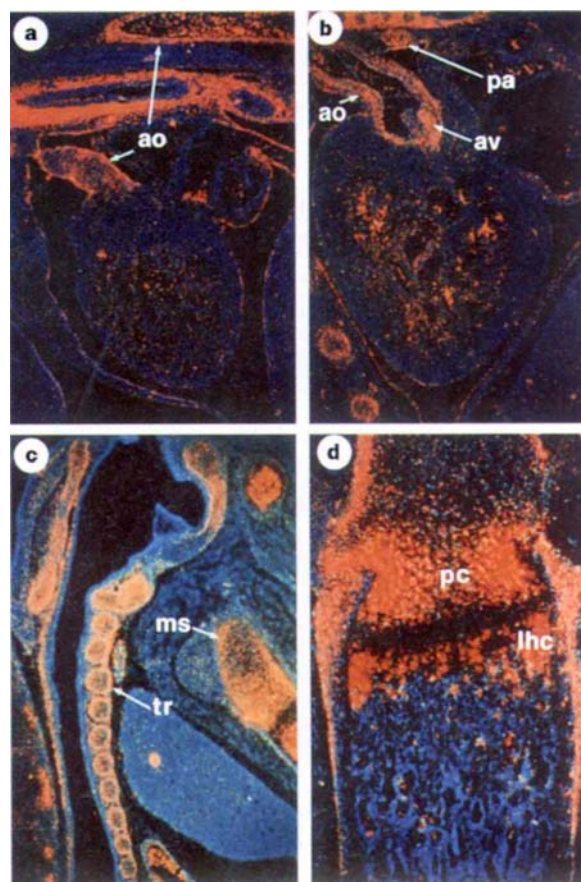


Figure 4 Analysis of *Mgp* expression during mouse development. *In situ* hybridization analysis in E12.5 (**a**) and E16.0 embryos (**b, c**), and 2-day-old pup (**d**). *Mgp* is highly expressed in the aorta (ao), pulmonary artery (pa), and in aortic (av) valve, but not in the myocardium (**a, b**). High levels of *Mgp* transcripts are also present in the trachea (tr) and chondrogenic zone of the manubrium sterni (ms) (**c**). In growth-plate cartilage, *Mgp* is highly expressed in proliferative chondrocytes (pc) and to a lesser extent in late hypertrophic chondrocytes (lhc) but not in early hypertrophic chondrocytes or osteoblasts (**d**).

diffraction, or were conventionally contrasted by brief staining with uranyl acetate and lead citrate. For analysis of demineralized specimens tissues were fixed in fresh 4% paraformaldehyde for 16 h, demineralized for 2 d in 10% formic acid, dehydrated, embedded in paraffin, sectioned at 5 μ m, and stained with haematoxylin/eosin.

In situ hybridization. Embryos and tissues were fixed in fresh 4% paraformaldehyde for 10–24 h, dehydrated, embedded in paraffin and sectioned at 8 μ m. For *in situ* hybridization, an antisense riboprobe corresponding to mouse *Mgp* cDNA³⁰ was used. Procedures have been described^{28,29}; hybridization and washes were carried out at 50°C and 64°C, respectively. Autoradiography, Hoechst 33258 staining, and photography were done as described²⁸.

Received 28 October; accepted 3 January 1997.

- Ross, R. *Nature* **362**, 801–809 (1993).
- Manjin, H. J. & Brandt, K. D. *Pathogenesis of Osteoarthritis* (eds Kelley, W. N., Harris, E. D., Ruddy, S. & Sledge, C. B.) 1469–1479 (Saunders, Philadelphia, 1989).
- Hauschka, P., Lian, J., Cole, D. & Gundberg, C. *Physiol. Rev.* **69**, 990–1047 (1989).
- Furie, B. & Furie, B. C. *Cell* **53**, 505–517 (1988).
- Ducy, P. *et al.* *Nature* **382**, 448–452 (1996).
- Dowd, P., Hershtine, R., Ham, S. W. & Naganathan, S. *Science* **269**, 1684–1691 (1995).
- Tanimura, A., McGregor, D. H. & Anderson, H. C. *J. Exp. Pathol.* **2**, 261–273 (1986).
- Solursh, M. *Cell and Matrix Interactions during Limb Chondrogenesis in Vitro* (ed. Trested, R. L.) 277–303 (Liss, New York, 1984).
- Anderson, H. C. *J. Cell Biol.* **35**, 81 (1967). **final page?**
- Tanimura, A., McGregor, D. H. & Anderson, H. C. *Proc. Soc. Exp. Biol. Med.* **172**, 173–177 (1983).
- Singleton, E. B. & Merten, D. F. *Pediatr. Radiol.* **1**, 2–7 (1973).
- Qiao, J. H. *et al.* *Arterioscler. Thromb.* **14**, 1480–1497 (1994).
- O'Brien, K. D. *et al.* *Circulation* **92**, 2163–2168 (1995).
- Reid, J. D. & Andersen, M. E. *Atherosclerosis* **101**, 213–224 (1993).
- Zhang, M. H., Reddick, R. L., Piedrahita, J. A. & Maeda, N. *Science* **258**, 468–471 (1992).
- Plump, A. S. *et al.* *Cell* **71**, 343–353 (1992).
- Nakashima, Y., Plump, A. S., Raines, W., Breslow, J. L. & Ross, R. *Arterioscler. Thromb.* **14**, 133–141 (1994).
- Breslow, J. L. *Science* **272**, 685–688 (1996).
- Ross, R. & Bornstein, P. *Sci. Am.* **224**, 44–52 (1971).
- Smithies, O. & Maeda, N. *Proc. Natl. Acad. Sci. USA* **92**, 5266–5272 (1995).
- Karaplis, A. C. *et al.* *Genes Dev.* **8**, 277–289 (1994).
- Vortkamp, A. *et al.* *Science* **273**, 613–622 (1996).
- Ramirez-Solis, R., Davis, A. & Bradley, A. *Meth. Enzymol.* **225**, 855–878 (1993).
- Soriano, P., Montgomery, C., Geske, R. & Bradley, A. *Cell* **64**, 693–702 (1991).
- McMahon, A. P. & Bradley, A. *Cell* **62**, 1073–1085 (1990).
- Kochhar, D. M. *Teratology* **7**, 289–298 (1973).
- McKee, M. D. & Nanci, A. *Microsc. Res. Tech.* **33**, 141–164 (1996).
- Sundin, O. H., Busse, H. G., Rogers, M. B., Gudas, L. J. & Eichele, G. *Development* **108**, 47–58 (1990).
- Wilkinson, D. G. *In Situ Hybridization: A Practical Approach* (ed. Wilkinson, D. G.) 257–263 (IRL, Oxford, 1992).
- Luo, G., D'Souza, R., Hogue, D. & Karsenty, G. *J. Bone Min. Res.* **10**, 325–334 (1995).

Acknowledgements. G.L., P.D. and M.D.M. contributed equally to this work. We thank I. Szwarc and J. Turgeon for technical assistance, and H. Bellen, A. Bradley, L. Etkin, W. Klein and M. Majeski for critically reading the manuscript. M.D.M. thanks the members of the Laboratory for Electron Microscopy (Montreal) for their contribution. P.D. is supported by a fellowship from the Arthritis Foundation. M.D.M. is supported by grants from the MRC of Canada and the FRSQ. G.K. is supported by grants from the NIH and the March of Dimes Foundation.

Correspondence and requests for materials to G.K. (e-mail: gkarsenty@molgen.mda.uth.tmc.edu).

Osteopetrosis in mice lacking haematopoietic transcription factor PU.1

M. M. Tondravi*, S. R. McKercher†, K. Anderson†, J. M. Erdmann*, M. Quiroz*, R. Maki† & S. L. Teitelbaum*

* Department of Pathology, Washington University School of Medicine, St Louis, Missouri 63110, USA

† The Burnham Institute, La Jolla, California 92037, USA

Osteoclasts are multinucleated cells and the principal resorptive cells of bone. Although osteoclasts are of myeloid origin¹, the role of haematopoietic transcription factors in osteoclastogenesis has not been explored. Here we show that messenger RNA for the myeloid- and B-cell-specific transcription factor PU.1 progressively increases as marrow macrophages assume the osteoclast phenotype *in vitro*. The association between PU.1 and osteoclast

differentiation was confirmed by demonstrating that PU.1 expression increased with the induction of osteoclastogenesis by either 1,25-dihydroxyvitamin D₃ or dexamethasone. Consistent with the participation of PU.1 in osteoclastogenesis, we found that the development of both osteoclasts and macrophages is arrested in PU.1-deficient mice. Reflecting the absence of osteoclasts, PU.1^{-/-} mice exhibit the classic hallmarks of osteopetrosis, a family of sclerotic bone diseases². These animals were rescued by marrow transplantation, with complete restoration of osteoclast and macrophage differentiation, verifying that the PU.1 lesion is intrinsic to haematopoietic cells. The absence of both osteoclasts and macrophages in PU.1-mutant animals suggests that the transcription factor regulates the initial stages of myeloid differentiation, and that its absence represents the earliest developmental osteopetrotic mutant yet described.

PU.1 (Spi-1, Sfpi-1) is an ETS-domain transcription factor essential for the development of myeloid and B-lymphoid cells^{3–5}. In myeloid cells, PU.1 is thought to regulate transcription of both *c-fms* and CD11b/CD18 (Mac1), proteins that are central to the osteoclast phenotype^{6–9}. The similarities between macrophages and osteoclasts, and the critical role PU.1 plays in generating myeloid cells (which are osteoclast progenitors), raised the possibility that failure to express PU.1 might preclude development of osteoclasts, leading to arrested bone resorption and osteopetrosis.

To test this idea, we first determined whether a temporal relationship exists between the appearance of osteoclasts and PU.1 expression *in vitro*. Osteoclasts were generated by co-culture of murine bone marrow mononuclear cells (BMMs) with the ST2 stromal line. Osteoclasts were identifiable by the fourth day of culture and reached a maximum by day 7–8. Stromal cells were then selectively removed by collagenase digestion, and RNA, derived from the residual population (osteoclasts) was probed with a PU.1 cDNA. As expected, macrophage colony stimulating factor (M-CSF)-dependent BMMs, namely macrophages, expressed a single 1.4-kilobase (kb) PU.1 mRNA (Fig. 1a). In contrast, PU.1 mRNA was expressed not only by BMMs in early co-culture, but its level increased 3-fold as the cells differentiated into osteoclasts (Fig. 1a). Thus, PU.1 is expressed throughout the entire temporal span of osteoclastogenesis. ST2 cells do not express PU.1 (data not shown).

Generation of murine osteoclasts by co-culture of BMMs and ST2 cells requires 1,25-dihydroxyvitamin D₃ (D₃) and dexamethasone. BMM/ST2 cells were co-cultured with an optimal dose of dexamethasone (10⁻⁸ M) and varying concentrations of D₃ (Fig. 1b). In cultures containing $\geq 10^{-9}$ M D₃, osteoclasts developed within 7 days. ST2 cells were removed and RNA derived from osteoclasts and their differentiating precursors was assayed for PU.1 expression by northern blot analysis (Fig. 1b). Mirroring the appearance of osteoclasts, PU.1 mRNA was upregulated only in cultures containing 10⁻⁸ and 10⁻⁹ M D₃. β_3 integrin mRNA, which serves as a molecular marker of osteoclast differentiation in this system, was also only expressed in cultures containing 10⁻⁸ and 10⁻⁹ M D₃ (Fig. 1b). The reciprocal experiment, in which osteoclastogenic cultures were established with an optimal dose of D₃ (10⁻⁸ M) and varying concentrations of dexamethasone, produced osteoclasts only in the presence of 10⁻⁸ and 10⁻⁹ M glucocorticoid, an event paralleled by β_3 integrin mRNA expression (Fig. 1c). Once again, PU.1 mRNA was most abundant in cultures containing mature osteoclasts, and declined, dose dependently, with decreasing dexamethasone. Thus, there exists a direct relationship between osteoclastogenesis, PU.1 mRNA expression and expression of a functional osteoclast integrin subunit, β_3 , whose promoter serves as a target of PU.1 (unpublished results).

The fact that PU.1 expression parallels osteoclastogenesis *in vitro* suggests an important role for PU.1 in osteoclast development. To assess the role of PU.1 *in vivo*, we deleted the transcription factor by engineering mice in which its coding sequence was disrupted by homologous recombination⁵. PU.1^{-/-} mice are born alive, but die

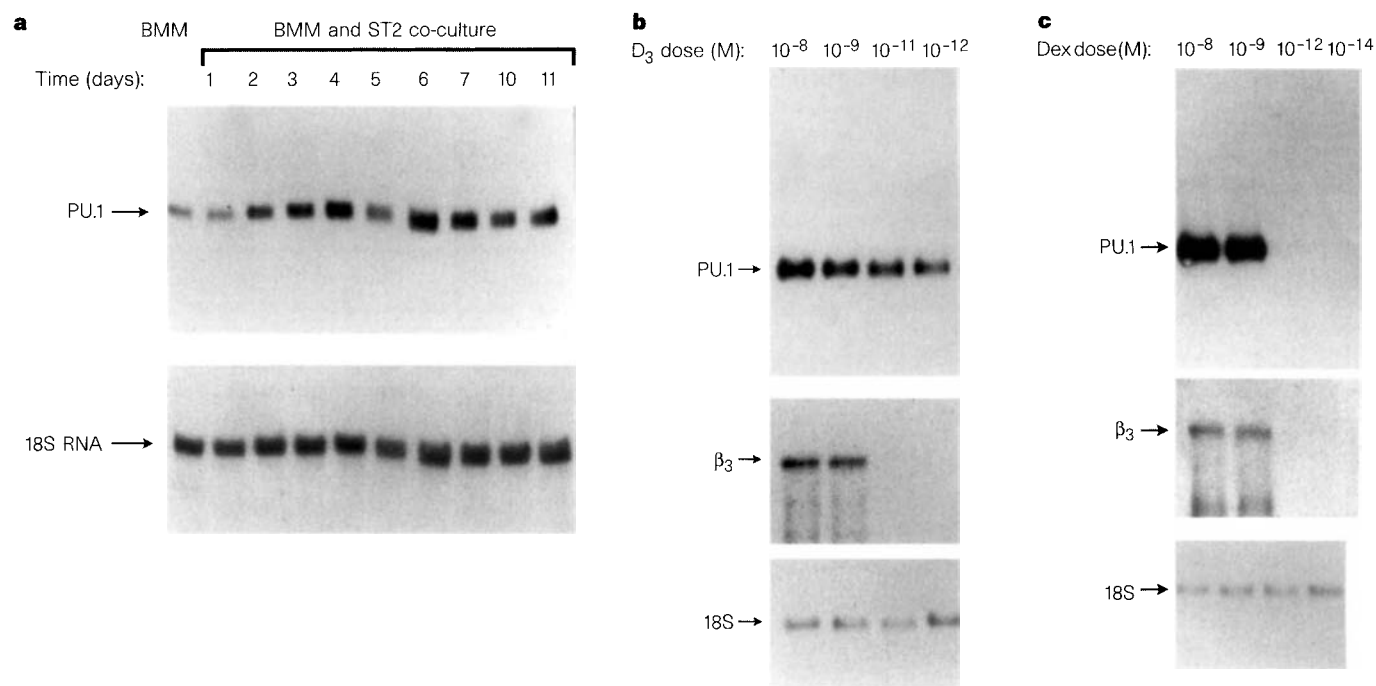


Figure 1 PU.1 expression parallels osteoclastogenesis. **a**, RNA from osteoclasts (lanes 7–11) or their precursors (1–6) or macrophages (BMM) was probed for PU.1 and 18S rRNA to normalize loading. **b**, RNA from osteoclasts, or their precursors, differentiated with 10^{-8} M dexamethasone and indicated concentrations of D_3 was probed for PU.1 (top), β_3 integrin (middle) and 18S rRNA. PU.1 and β_3 integrin mRNA levels are undetectable at concentrations of D_3 below 10^{-9} M, which also

fail to support osteoclastogenesis. **c**, RNA from osteoclasts or their precursors, differentiated with 10^{-8} M D_3 and indicated concentrations of dexamethasone, was probed for PU.1 (top), β_3 integrin (middle), and 18S rRNA (bottom). PU.1 mRNA mirrors dexamethasone levels and β_3 integrin mRNA is undetectable at steroid concentrations below 10^{-9} M, the minimum supporting osteoclastogenesis.

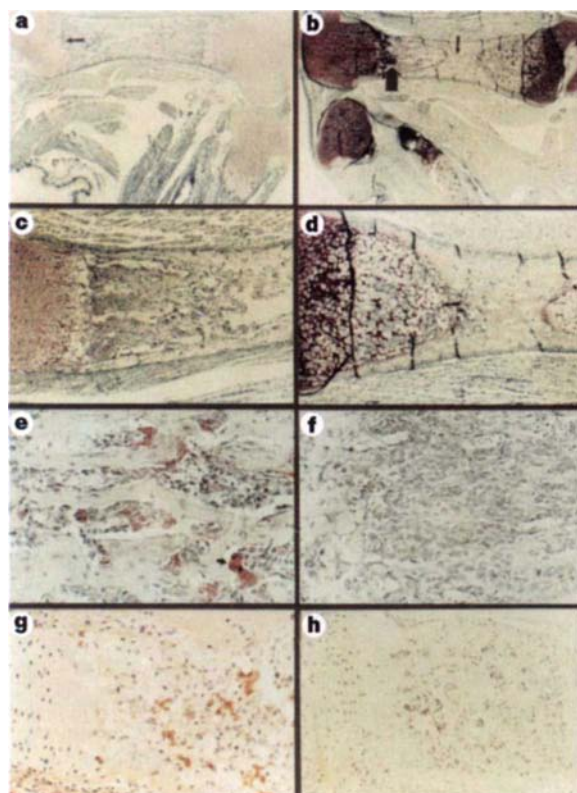


Figure 2 PU.1 homozygous mutant mice develop osteopetrosis. Histological sections of tibia from a one-day-old PU.1 heterozygote (**a**, **c**, **e**, **g**) and a homozygous mutant littermate (**b**, **d**, **f**, **h**) stained with toluidine blue to identify cartilage (**a–d**), for TRAP activity to delineate osteoclasts (**e** and **f**), or with F4/80 monoclonal antibody to detect macrophages (**g** and **h**). Heterozygous bone is structurally normal (**a** and **c**), with distinct demarcation between growth plate and primary spongiosa (**a**, arrow), a feature absent in the mutant (**b**, bold arrow). At higher magnification (**c**, **d**) metachromatic cartilaginous 'bars' are evident deep within the metaphysis of the homozygous mutant (**d**, arrow). Furthermore, in contrast to heterozygote animals in which primary appositional bone has been remodelled into normal trabeculae, no such remodelling has occurred in the PU.1^{-/-} mutant as primitive, broad trabeculae persist (small arrow in **b**). The PU.1^{-/-} mutant lacks TRAP⁺ (red) osteoclasts (**f**), which are abundant in the heterozygous mouse (**e**, arrow). Marrow of a PU.1 heterozygote contains abundant brown-stained, F4/80⁺ macrophages (**g**), absent in the homozygote mutant (**h**).

Figure 3 Bone histology of a $PU.1^{-/-}$ mouse rescued by marrow transplantation. **a**, marrow of a rescued $PU.1^{-/-}$ mouse probed with F4/80, showing normal abundance of brown-stained macrophages (compare with Fig. 2g-h). **b**, Marrow of the rescued mouse, similar to the control, contains abundant TRAP⁺ (red) osteoclasts (compare with Fig. 2e, f). **c**, Tibia of a rescued mouse, stained with toluidine blue demonstrates nearly complete resolution of the osteopetrotic phenotype (compare with Fig. 2a-d). The demarcation between the growth plate and primary spongiosa in the rescued animal is similar to that of the heterozygote and is more distinct than that of the mutant. In addition, virtually all cartilaginous bars (blue) present in the mutant are remodelled in the rescued mouse.

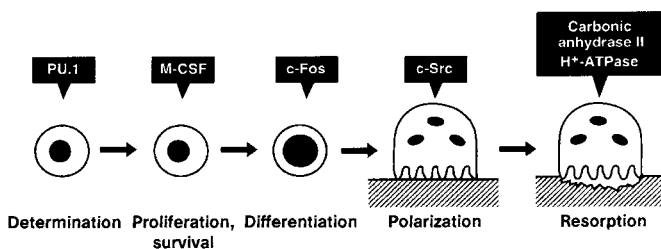
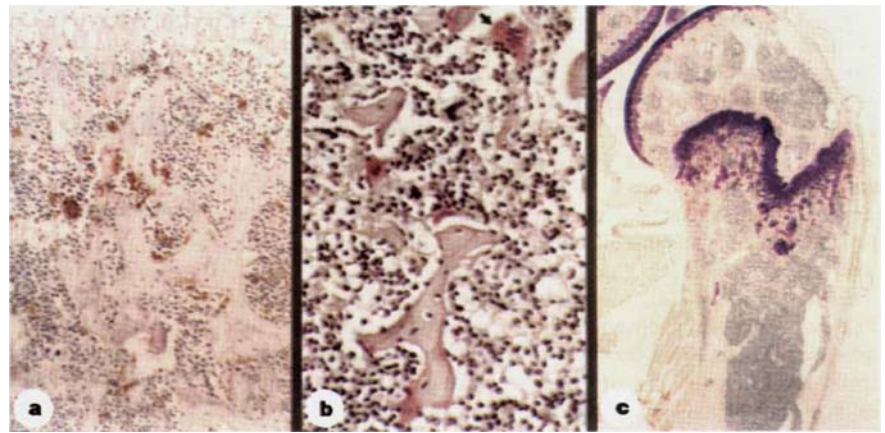


Figure 4 $PU.1^{-/-}$ exerts its effect on osteoclast development earlier than any other known osteopetrotic mutation. Whereas $PU.1$ and $c-fos$ both affect osteoclastogenesis, $PU.1$ impacts on an earlier step in the developmental pathway that affects both macrophages and osteoclasts. The mutations of $c-src$, carbonic anhydrase II, and H^+ -ATPase affect the function of the mature osteoclast. The absence of M-CSF in the op/op mouse is a mutation primarily affecting stromal cells and osteoblasts, both of which normally express the cytokine, and thus indirectly targets osteoclasts or their progenitors.

of septicaemia within 24–48 hours. The skeletons of these animals exhibited pathognomonic histological features of osteopetrosis (Fig. 2). Metachromatic staining demonstrated a sharp demarcation between the growth plate (physis) and metaphysis in long bones of one-day-old $PU.1^{+/+}$ phenotypically normal mice, reflecting intact osteoclastic resorption of primary spongiosa (Fig. 2a, c). This anatomical demarcation was absent in bones of $PU.1^{-/-}$ littermates. Loss of clear physal/metaphyseal delineation in the mutant reflects the failure to resorb primary spongiosa resulting in the microscopic hallmark of all forms of osteopetrosis, that is, persistence of cartilaginous 'bars' deep in metaphyseal bone (Fig. 2b, d). In contrast to $PU.1^{-/-}$ animals, in which primary appositional bone was modelled into normal trabeculae, no such modelling occurred in homozygous mutants, as primitive, broad cancellum persisted (Fig. 2c, d). Consistent with the failure to resorb cartilaginous bars and appositional bone, the $PU.1$ homozygous mutants were devoid of osteoclasts, which were abundant in the heterozygote (Fig. 2e, f). These osteopetrotic characteristics were observed in all long bones as well as the vertebrae of $PU.1$ homozygous mutant mice regardless of age. Some of the $PU.1^{-/-}$ animals were kept alive for as long as two weeks by antibiotic treatment. Examination of their skulls by X-ray illustrated another hallmark of osteopetrosis, namely that the incisors of the mutants failed to erupt, whereas those of the wild-type littermates descended (data not shown).

Osteoclasts are myeloid-derived cells, closely related to macrophages. $PU.1^{-/-}$ mice are not only devoid of osteoclasts, but also lack macrophages within the marrow (Fig. 2g, h), lungs and liver⁸, as indicated by the lack of F4/80 staining of cells within these tissues. Thus, deletion of $PU.1$ prevents osteoclast and macrophage differentiation.

The fact that $PU.1$ is expressed in haematopoietic cells indicated that the osteopetrotic phenotype should be rescuable by marrow transplantation. Following intraperitoneal injection of 10^7 marrow cells recovered from four-week-old $PU.1^{+/+}$ or $PU.1^{+/-}$ mice, into newborn $PU.1^{-/-}$ mutants, the majority of transplanted animals survived for at least six months and were grossly indistinguishable from wild-type littermates. All aspects of the osteopetrotic phenotype were also rescued, including incisor eruption (data not shown),

modelling of the primary appositional bone into normal trabeculae, and a clear demarcation between growth plate and primary spongiosa (compare Figs 2, 3). Moreover, nearly all metaphyseal cartilaginous bars (Fig. 2b, d) were resorbed in the rescued animals (Fig. 3). The persistence of occasional bars undoubtedly reflected the osteopetrotic history of the rescued mice before transplantation. Finally, the marrow of successfully transplanted animals contained abundant F4/80-positive macrophages and multinucleated osteoclasts positive for tartrate-resistant acid phosphatase (TRAP) (Fig. 3). Rescue of $PU.1^{-/-}$ mutants by marrow transplantation confirms that failure to generate macrophages and osteoclasts in the $PU.1$ homozygous mutant is due to a cell-autonomous defect in a common haematopoietic progenitor(s) and not an altered haematopoietic stromal environment. In addition, the $PU.1$ mutation is consistent genetically with a common haematopoietic lineage of these two cell types.

Mutations generating the osteopetrotic phenotype provide insights into the molecular mechanisms of bone resorption. Previously described mutants include those arresting osteoclast development from myeloid precursors, such as the op/op and the $c-fos^{-/-}$ mutations in mice, or those impairing resorptive activity of the mature cell, which include the $c-src^{-/-}$ mouse, as well as the carbonic anhydrase II and the H^+ -ATPase mutations in man^{10–18}. The op/op mouse, which does not express functional M-CSF, represents a defect in stromal and osteoblastic cells essential for osteoclastogenesis, whereas the molecular defects in the $PU.1^{-/-}$ and the $c-fos^{-/-}$ mice lie within osteoclast precursors *per se*^{10–15}. The op/op mouse spontaneously recovers with age, indicating that M-CSF, although it is permissive for osteoclastogenesis, is not requisite¹³. In $c-fos$ mutants, the number of marrow macrophages supersedes that in wild-type mice, suggesting the proto-oncogene may promote differentiation of a bipotential macrophage/osteoclast precursor into bone resorptive polykaryons¹⁴. Because the $PU.1^{-/-}$ mouse lacks both osteoclasts and macrophages, it represents the earliest developmental form of osteopetrosis yet described (Fig. 4).

Another $PU.1$ -knockout mouse^{4,5} generated by Scott *et al.* shares some similarities with our mouse, such as a lack of macrophages and B cells. The principal differences between the two mice are that (1)

our mouse survives to birth and beyond, whereas that of Scott *et al.*⁴ dies by embryonic day 18 (E18); and (2) the skeletal histology of the mouse generated by Scott *et al.*⁴ has not been described. Examination of our mice at E18 demonstrated a lack of osteoclasts. Because the animal described by Scott *et al.*⁴ lacks multipotential myeloid progenitors, it too probably fails to generate osteoclasts. Although a number of potential causes for the difference in survival of the two mice have been suggested⁵, we have excluded the possibility that read-through transcription generates a partial PU.1 protein in our animal (data not shown).

The PU.1 osteopetrotic mutation represents a cell-autonomous defect that results in a complete absence of both macrophages and osteoclasts. These mice thus provide a unique opportunity to identify the intermediates in osteoclast lineage differentiation *in vivo*. For example, wild-type myeloid precursors can be differentiated to definable stages of osteoclast development *in vitro* and introduced into the marrow of PU.1^{-/-} animals. The differentiation potential of these lineage-committed cells to osteoclasts and macrophages can then be determined *in vivo*. □

Methods

Osteoclastogenic cultures. *In vitro* differentiation of osteoclasts and bone marrow macrophages has been described^{19,20}.

Northern blot hybridization. ST2 cells were removed from osteoclastogenic cultures by digestion with 10 µg ml⁻¹ collagenase (type 1, Sigma)²⁰. RNA from osteoclasts and their precursors was isolated using Trizol (BRL). RNA samples (7 µg) were run on 1.1% agarose-formaldehyde gels, blotted and hybridized with random-prime labelled PU.1 or β₃ cDNAs, or an end-labelled oligonucleotide, 5'-CATGGTAGGCACGCGACTACCAT-3', which hybridizes to the 18S ribosomal RNA (rRNA) probe²¹.

Maintenance of PU.1 transgenic mice. Homozygous PU.1 mutant mice, bred from heterozygote parents, were genotyped by PCR using tail DNA as template⁵. In some cases, homozygous mutant PU.1 newborns were injected subcutaneously twice per day with Baytril (enrofloxacin, 2.5 mg kg⁻¹).

Bone marrow transplantation experiments. Femurs and tibiae of four-week-old male PU.1^{+/+} or PU.1^{+/-} mice were flushed, the marrow mononuclear cells isolated by centrifugation on a 1.077 g ml⁻¹ density Ficoll-Hypaque cushion and resuspended in PBS at 10⁸ cells per ml. Each newborn PU.1^{-/-} mouse was injected intraperitoneally with ~10⁷ marrow mononuclear cells.

Histology. All tissues were processed for histology and stained for TRAP activity as described²². Metachromasia was assessed by toluidine-blue staining (0.7%). Macrophages were identified by immunostaining tissue sections with 10 µg ml⁻¹ of F4/80 (Serotec) using a Vectastain-HRP kit (Vector Labs).

Received 13 November 1996; accepted 13 January 1997.

- Teitelbaum, S. L., Tondravai, M. M. & Ross, F. P. in *Osteoporosis* 61–94 (Academic, San Diego, 1996).
- Felix, R., Hofstetter, W. & Cecchini, M. G. *Eur. J. Endocrinol.* **134**, 143–156 (1996).
- Klemsz, M. J., McKercher, S. R., Celada, A., VanBeveren, C. & Maki, R. A. *Cell* **61**, 113–124 (1990).
- Scott, E. W., Simon, M. C., Anastasi, J. & Singh, H. *Science* **265**, 1573–1577 (1994).
- McKercher, S. R. *et al.* *EMBO J.* **15**, 5647–5658 (1996).
- Zhang, D.-E. *et al.* *Molec. Cell. Biol.* **14**, 8085–8095 (1994).
- Pahl, H. L. *et al.* *J. Biol. Chem.* **268**, 5014–5020 (1993).
- Rosmarin, A. G., Caprio, D., Levy, R. & Simkevich, C. *Proc. Natl Acad. Sci. USA* **92**, 801–805 (1995).
- Duong, L. T., Tanaka, H., Wesolowski, G. & Rodan, G. A. *J. Bone Miner. Res.* **8**, S378 (1993).
- Yoshida, H. *et al.* *Nature* **345**, 442–443 (1990).
- Takahashi, N. *et al.* *Endocrinology* **128**, 1792–1796 (1991).
- Kodama, H., Nose, M., Nilda, S. & Yamasaki, A. *J. Exp. Med.* **173**, 1291–1294 (1991).
- Begg, S. J. *et al.* *J. Exp. Med.* **177**, 237–242 (1993).
- Grigoriadis, A. E. *et al.* *Science* **266**, 443–448 (1994).
- Wang, Z. Q. *et al.* *Nature* **360**, 741–745 (1992).
- Felix, R., Cecchini, M. G. & Fleisch, H. *Endocrinology* **127**, 2592–2594 (1993).
- Yamamoto, T. *et al.* *J. Clin. Invest.* **91**, 362–367 (1993).
- Sly, W. S., Hewett-Emmett, D., Whyte, M. P., Yu, Y. & Tashian, R. E. *Proc. Natl. Acad. Sci. USA* **80**, 2752–2756 (1983).
- Clohisy, D. R., Bar-Shavit, Z., Chappel, J. & Teitelbaum, S. L. *J. Biol. Chem.* **262**, 15922–15929 (1987).
- Shioi, A., Ross, F. P. & Teitelbaum, S. L. *Calcif. Tissue Int.* **55**, 387–394 (1994).
- Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning. A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, Plainville, New York, 1989).
- Lacey, D. L. *et al.* *Endocrinology* **136**, 2367–2376 (1995).

Acknowledgements. We thank T. Tolley for assistance with histological preparation, and P. Ross and P. Powell for criticism. Supported by grants from the Shriners Hospital (S.L.T.), the Barnes-Jewish Hospital Foundation (M.M.T.) and the NIH (M.M.T., S.L.T., R.M.).

Correspondence and requests for materials should be addressed to M.M.T. (e-mail: tondravm@medicine.wustl.edu).

Homeosis and intestinal tumours in *Cdx2* mutant mice

K. Chawengsaksophak*, R. James†, V. E. Hammond*, F. Köntgen‡ & F. Beck*

* Howard Florey Institute of Experimental Physiology and Medicine, University of Melbourne, Parkville, Victoria 3052, Australia

† Department of Medicine, The Queen Elizabeth Hospital, Woodville, South Australia 5011, Australia

‡ Walter and Eliza Hall Institute of Medical Research, Royal Melbourne Hospital, Victoria 3050, Australia

In *Drosophila*, disturbing the expression of the homeobox gene *caudal* causes a severe disruption in body segmentation and global body patterning¹. There are three mouse homologues of *Drosophila caudal*: *Cdx1* (ref. 2), *Cdx2* (ref. 3) and *Cdx4* (ref. 4). We have generated a null mutation of murine *Cdx2* by homologous recombination. *Cdx2* homozygote null mutants die between 3.5 and 5.5 days post coitum (d.p.c.). *Cdx2* heterozygote mutants exhibit a variable phenotype, with many showing tail abnormalities or stunted growth. Skeletal analysis demonstrates a homeotic shift of vertebrae and compatible malformations of the ribs. Within the first three months of life, 90% of *Cdx2* heterozygotes develop multiple intestinal adenomatous polyps, particularly in the proximal colon. These polyps occasionally contain areas of true metaplasia. In contrast to the surrounding intestinal epithelium, the neoplastic cells do not express *Cdx2* from the remaining allele. These results suggest that *Cdx2* mutation is the primary event in the genesis of some intestinal tumours.

In the mouse, *Cdx2* is expressed extra-embryonically at 3.5 days post coitum (d.p.c.) in the trophectoderm and later in some trophectodermally derived placental tissues. Embryonic expression begins at 8.5 d.p.c. in the posterior gut, the tailbud, the posterior part of the neural tube, and the unsegmented paraxial mesoderm before the development of somites⁵. In the later embryo and adult, *Cdx2* expression is confined to the intestinal epithelium, with expression being highest in the proximal colon⁶. We generated embryonic stem (ES) cell lines that have one non-functional allele of *Cdx2* by conventional gene targeting⁷ (Fig. 1). Nine *Cdx2* knock-out cell lines were isolated, with a targeting efficiency of approximately 1%. Three of these were selected for the generation of chimaeric mice by blastocyst injection; two (215B and 305C) produced chimaeric mice that transmitted the mutation to their progeny. Heterozygote offspring were viable and fertile.

Genotyping of postimplantation embryos and postnatal animals showed that offspring were either heterozygous or wild type (Table 1). Immediately before implantation (3.5 d.p.c.), homozygous negative blastocysts were demonstrated by the polymerase chain reaction (PCR) to have the expected mendelian 1:2:1 distribution ratio (Table 1). Thus homozygous null mutants do not survive the peri-implantation period. The earliest expression of *Cdx2* during development occurs in the trophectoderm at 3.5 days (ref. 5), and this may explain why null mutant blastocysts fail to implant successfully.

Cdx2 heterozygous abnormalities were the same for both of the cell lines used in their production. A shortened or kinky tail was found in 45 of 75 (60%) heterozygotes derived from cell line 215B, and 43 of 86 (50%) mice derived from 305C. For cell line 215B, 5 of 18 (28%), and for cell line 305C, 5 of 9 (56%), heterozygote neonate weights were more than 2 standard deviations below that of their wild-type littermates (Fig. 2). *Cdx2* is normally expressed in the spongy cytotrophoblastic cells of the placenta (descendants of the trophectoderm). Disturbance of their growth would result in placental insufficiency leading to fetal death or growth retardation.

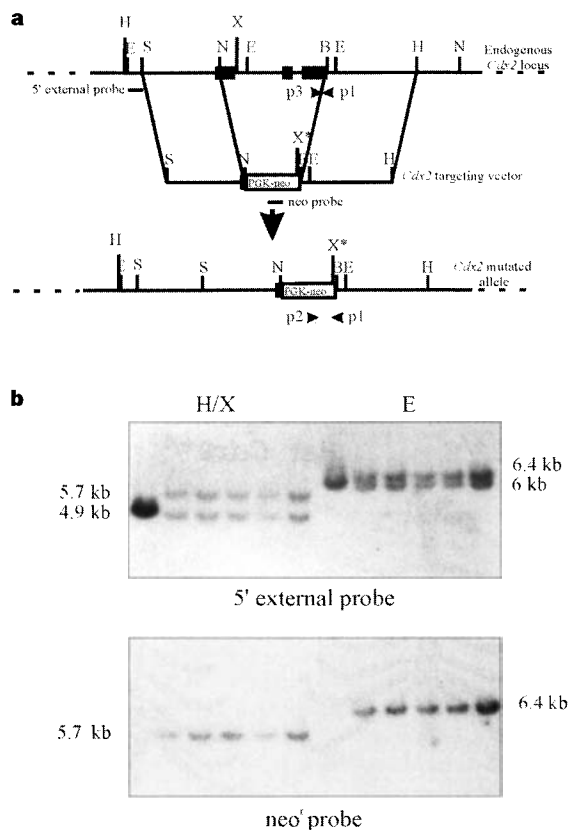


Figure 1 Generation of *Cdx2* mutant mice. **a**, Schematic representation of the endogenous mouse *Cdx2* locus (top), *Cdx2* targeting vector (middle), and mutant allele (bottom). *Cdx2* coding exons are shown in black boxes and the PGK-*neo*^r cassette is indicated by the grey box. The 5' external and internal *neo*^r probes used for genomic Southern analysis of ES cell colonies (see below) are shown in black lines. Arrows indicate the primers (p1, p2 and p3) used for genotyping progeny of heterozygote matings by PCR (see Table 1). Restriction enzyme sites: B, *Bam*HI; E, *Eco*RI; H, *Hind*III; N, *Not*I; S, *Sac*I; X, *Xba*I. The *Xba*I restriction site (X*) was introduced during construction of the *Cdx2* targeting vector. The bacterial *neo*^r selection cassette, driven by a phosphoglyceratekinase promoter (PGK-*neo*^r), together with 4-kb 5' and 5.4-kb 3' flanking sequences, served to replace 5.9 kb of the native gene. **b**, Southern analysis of five *Cdx2* targeting cell lines. The left lane shows a wild-type W 9.5 cell line, and the rest show *Cdx2* targeted cell lines. In both panels, samples on the left half of the blot were digested with *Hind*III and *Xba*I, those on the right with *Eco*RI. Top, the blot was hybridized with the 5' external probe; bottom, the same filter was probed with a fragment from the *neo*^r gene. **c**, PCR assay used to genotype 3.5-d.p.c. progeny of heterozygote intercrosses showing three possible genotypes: wild-type (+/+), heterozygote (+/-), and homozygote (-/-).

Definitive conclusions concerning placental function must await detailed morphometric analysis.

A survey of sagittal sections from heterozygote embryos taken at 10.5, 11.5, 13.5, 14.5 and 16.5 d.p.c. showed normal histogenesis. No gross anatomical disturbances were observed except for one fetus at 11.5 d.p.c. in which widespread cell necrosis and tissue disruption indicated early resorption.

Skeletal analysis of heterozygotes showed an anterior homeotic shift of the cervical and thoracic spine in 15 of 16 animals examined (compared with 10 normal wild-type mice) (Fig. 3a and see Supplementary Information). Rib abnormalities, in keeping with this phenotype, were also apparent (Fig. 3b). These results indicate that the *Cdx2* gene is haplo-insufficient, and that loss of function leads to an anterior homeotic shift. The anterior shift of vertebrae in *Cdx2* heterozygotes is similar to that described for *Cdx1* homozygote null mutants⁸.

In adult mice, *Cdx2* expression is restricted to the intestine, where transcript levels vary quantitatively along the rostrocaudal axis⁶. Examination of the gut of *Cdx2*^{+/-} mice between 12 and 28 weeks of age revealed the presence of multiple adenomatous polyps in 20 of 22 animals, but none in age-matched wild-type littermates. Tumours (1–10 per animal) were found predominantly in the proximal colon and occasionally in the small intestine, a regional distribution that corresponds to the site of maximal *Cdx2* expression⁶. Tumour histology was the same for cell lines 215B and 305C.

The largest tumours (4 × 8 mm) were pedunculated tubulovillous colonic adenomata (Fig. 4a, b). Microscopic examination showed some hyperplastic areas without evidence of dysplasia, but the predominant picture was of grossly altered crypt architecture with obvious dysplasia and occasional giant cells (Fig. 4b, c). Many mitoses were present in cells throughout the tumours. The crypt architecture in regions adjacent to the tumours was sometimes deformed as a result of local tissue distortion. The epithelial cells in these crypts were not dysplastic, and in common with the

normal-looking gut epithelium, such areas retained *Cdx2* staining on immunocytochemical examination; by contrast, epithelial cells in the tumours ceased to express *Cdx2* (Fig. 4d). We also stained tumours from *Min*^{+/-} mice (see below) for *Cdx2* and found a similar extinction of expression. These findings agree closely with previous data from studies in which *Cdx2* expression was examined in a range of human and rat colonic neoplasms⁹. In the small intestine, frank tumours (Fig. 4e) were less frequent but there were extensive regions in which the crypts and villi had an abnormal architecture.

We examined tumour tissue and adjacent gut for loss of heterozygosity. Using Southern blot analysis of DNA from dissected tumour tissues, we were unable to demonstrate loss of heterozygosity in any of 18 tumours studied (see Supplementary Information). Loss of *Cdx2* expression from the remaining *Cdx2* allele in tumour



Figure 2 Four heterozygote mice at 3 weeks of age (left) compared to a wild-type animal (right). The phenotypic effect varies in severity.

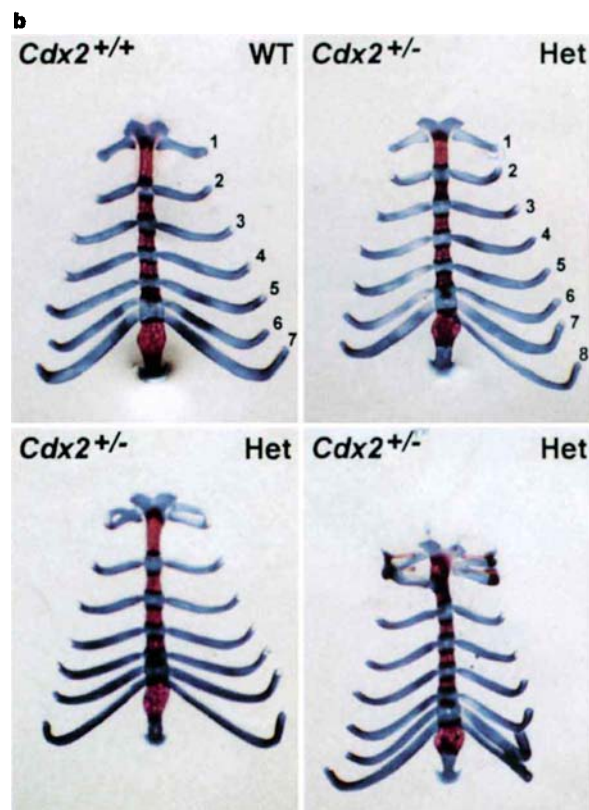
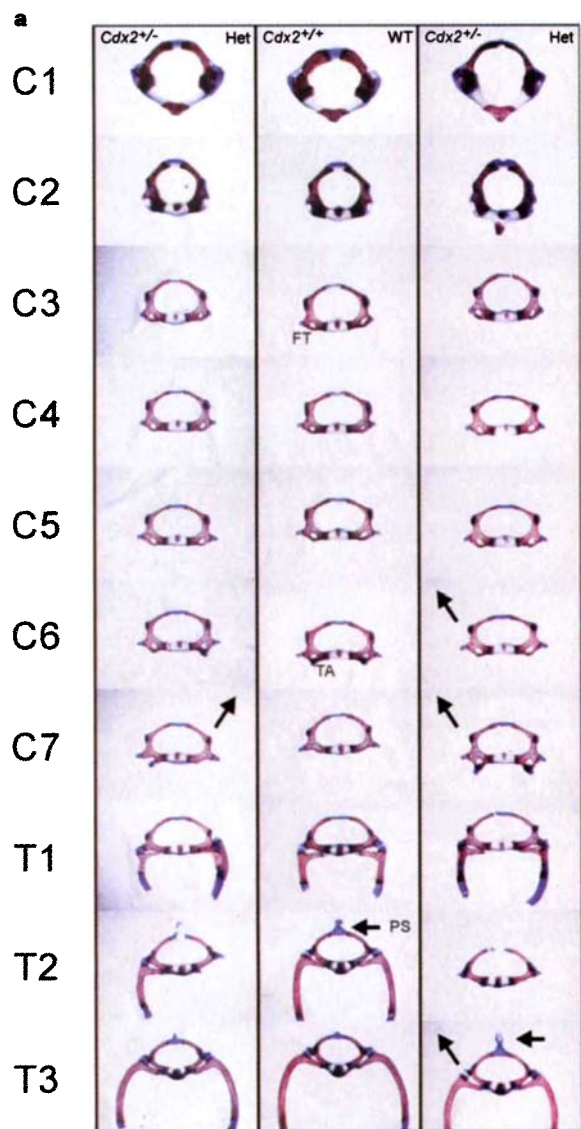


Figure 3 a, Skeletal preparations showing the morphology of the cervical and upper thoracic vertebrae in *Cdx2*^{+/-} heterozygotes (Het) and a wild-type (WT) littermate. The vertebral number is shown on the left. C6, C7 and T3 from heterozygous animals show morphological features characteristic of the immediately cranial vertebrae in the WT littermate control. This is seen in expression of anterior tubercles and foramina transversariae in heterozygote C7, absence of these tubercles in heterozygote C6, and the presence of a prominent spinous process in heterozygote T3 (arrows). FT, Foramen transversum; TA, anterior tubercle; PS, spinous process. **b**, Malformations of the ribs. Sternum and ribs from three *Cdx2*^{+/-} heterozygotes and one WT littermate showing an anterior homeotic shift in the heterozygotes. The heterozygote in the top right has an eighth rib attached to the sternum on the left side. The heterozygote in the bottom left has the second rib attached to the sternum at the top of the manubrium together with the first rib. The latter is incomplete and partly attached to the second rib (as is often seen in a cervical rib). The eighth rib is attached to the sternum bilaterally. The heterozygote in the bottom right has a similar arrangement on the left side whereas the second rib has retained its normal sternal attachment on the right.

tissue could be due to mutations in an important regulatory sequence, but is more likely to be a consequence of the phenotypic changes that occur during neoplastic transformation.

We also observed several areas of metaplasia within tumours (Fig. 4f). These took the form of keratinized epithelia that are very similar to oesophageal epithelium. It has been proposed that metaplasia, particularly in the context of renewing epithelia, might represent the mammalian equivalent of the more familiar homeotic transformations seen in lower species¹⁰. The idea that the resulting disruption of normal topographical tissue relationships could itself effect the induction or progression of neoplasia could be particularly relevant in this situation¹¹.

Of the phenotypic features observed, the most striking and least expected were the intestinal neoplasias and metaplasias. The neoplasias have some histological similarities to those of the *Min*^{+/-} mouse, which is caused by a mutation of the *Apc* gene¹². As in the *Min*^{-/-} mouse, homozygous null mutants do not develop to term, but death in the *Min*^{-/-} mutants is caused by failure of ectodermal

development following implantation, while trophoblastic cells are easily demonstrable at *Min*^{-/-} implantation sites¹³. No such appearances are manifest in *Cdx2*^{-/-} conceptuses. Furthermore, *Min*^{+/-} mice do not develop metaplasia, and show a regional incidence of tumours that differs considerably from that reported here. An initially shared molecular pathway in the genesis of *Min*^{+/-} and *Cdx2*^{+/-} tumour types therefore seems unlikely. The genes lie on different chromosomes, so mutation of the *Cdx2* gene, which lies on chromosome 5 (ref. 14), cannot have physically affected the *Apc* locus, which lies on chromosome 18 (ref. 15).

There is a growing body of evidence suggesting a link between homeobox gene expression and cellular proliferation¹⁶. However, we are not aware of any example of the loss of homeobox gene function resulting in a tumorigenic phenotype. How the loss of one *Cdx2* allele predisposes mice to develop intestinal tumours is not yet known. Our inability to demonstrate loss of heterozygosity suggests that *Cdx2*^{+/-} mice might be an appropriate model for investigating the hypothesis¹¹ that some forms of cancer may represent a

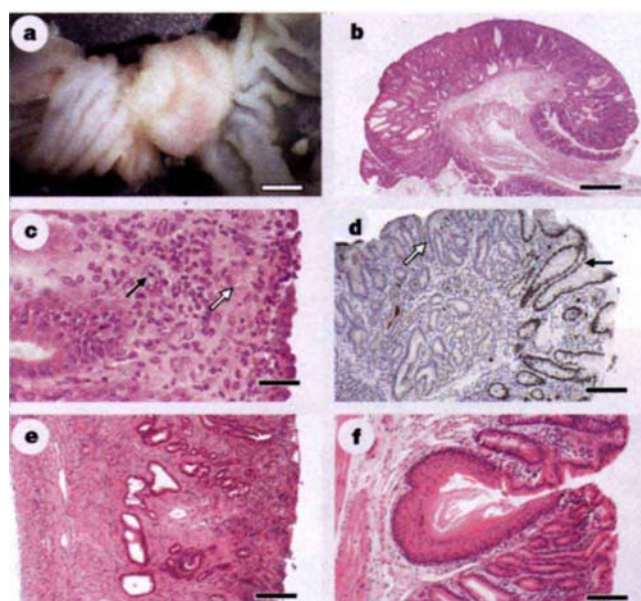


Figure 4 Tumours of the large and small intestine in *Cdx2*^{+/-} mice. **a**, Gross appearance of a tumour at junction of caecum and proximal colon. Scale bar, 1.4 mm. **b**, Low-power section of pedunculated adenoma from proximal colon showing abnormal crypt architecture. Haematoxylin and eosin (H&E) stain. Scale bar, 1,000 μ m. **c**, Giant cells (black arrow) and dysplastic epithelial cells (open arrow) from a tubulovillous adenoma of the proximal colon. H&E stain. Scale bar, 60 μ m. **d**, Absence of *Cdx2* staining (open arrow) in the epithelial cells of a tubulovillous adenoma from the proximal colon with clearly visible staining of adjacent mucous secreting crypts (black arrow). Scale bar, 170 μ m. **e**, Section through a sessile tumour of the small intestine. H&E stain. Scale bar, 240 μ m. **f**, Metaplastic region of keratinizing epithelium within a large bowel adenoma. H&E stain. Scale bar, 270 μ m.

Table 1 Genotyping of offspring from *Cdx2* heterozygote intercrosses

Age	Litters	Resorptions	Embryos typed	+/+	+/-	-/-
3.5 d.p.c.	11		91	20	52	19
4.5 d.p.c.*	6		46		45	1
5.5 d.p.c.*	7		47		47	0
6.5–10.5 d.p.c.	8	3	56	19	37	0
11.5–18.5 d.p.c.	8	6	48	20	28	0
New born	9		55	22	33	0
28 days	32		181	82	99	0

* Genotyping was determined by immunostaining rather than by PCR. In these animals, it is not possible to distinguish wild-type embryos from heterozygotes. (See Supplementary Information.)

disordered form of normal development. An alteration in the balance of factors within the regulatory hierarchies that determine normal cellular phenotype may manifest itself as neoplasia or metaplasia.

Human *CDX2* has been localized to chromosome 13q12.3 (ref. 17). We are currently investigating several samples of human colon cancer for evidence of allelic loss or mutation of this region. □

Methods

***Cdx2* targeting.** A 129Sv mouse genomic library was screened with fragments from both 5' and 3' regions of the previously characterized murine (BALB/c) *Cdx2* gene⁶. The final targeting vector consists of a 5' *Cdx2* 4-kilobase *SacI*–*NotI* fragment, followed by a PGK-*neo*^r from pPNT¹⁸ and a 3' *Cdx2* 5.8-kb, *Bam*HI–*Hind*III fragment. The targeting vector was linearized with *Hind*III and electroporated into a W 9.5 ES cell line¹⁹. After 11 days of G418 selection, 912 drug-resistant colonies were screened by Southern blot analysis. Nine clones yielded the expected restriction fragments. Two were selected to produce chimaeric mice by micro-injection into C57BL/6J × C57BL/10ScSn

host blastocysts. Male chimaeric mice were mated with C57BL/6J female mice to obtain germline transmission of the mutant allele. PCR was used for genotyping the offspring of *Cdx2* heterozygote crosses. Briefly, blastocysts were flushed from the uterus, and conceptuses older than 5.5 d.p.c. were dissected free of maternal tissues. To isolate DNA for PCR, both embryonic and neonatal tissues were lysed in PCR lysis buffer²⁰ with proteinase K (100 μ g ml⁻¹) at 56 °C. The PCR mixture contained the three primers shown in Fig. 1a: p1 (5'-TAAAGTCACTGTGTTCGGAATCC-3'), p2 (5'-ATATTGCTGAAGAGCTTGGCGGC-3') and p3 (5'-GGGACTATTCAAAG-TACAGGAG-3'). The reaction conditions have been described²¹. A 636-bp amplification product is generated from the mutant allele between the *neo*^r primer p2 and the *Cdx2* endogenous primer p1; a 443-bp product is amplified from the wild-type allele between p1 and p3.

Southern blot analysis for loss of heterozygosity. Tumour tissue was removed under a dissecting microscope. The normal surrounding tissue within 1 cm of each tumour was used as control. Tissues were digested in lysis buffer²² containing proteinase K (200 μ g ml⁻¹) at 56 °C. DNA was digested with *Eco*RI and hybridized with a 700-bp *Nco*I fragment that lies immediately 3' to the third coding exon of the *Cdx2* gene.

Histology and *in situ* detection of protein expression. Embryos were processed and stained with *Cdx2* antibody⁶. Sections were stained with haematoxylin and eosin for histological analysis. All sections studied were 5 μ m thick. Skeletal preparations were stained with alcian blue and with alizarin red S²³.

Received 30 September; accepted 31 December 1996.

- Macdonald, P. M. & Struhl, G. A molecular gradient in early *Drosophila* embryos and its role in specifying body pattern *Nature* **324**, 537–545 (1986).
- Duprey, P. *et al.* A mouse gene homologous to the *Drosophila* gene *caudal* is expressed in epithelial cells from the embryonic intestine *Genes Dev.* **2**, 1647–1654 (1988).
- James, R. & Kazenwadel, J. Homeobox gene expression in the intestinal epithelium of adult mice *J. Biol. Chem.* **266**, 3246–3251 (1991).
- Gamer, L. W. & Wright, C. V. E. Murine *cdx-4* bears striking similarities to the *Drosophila* *caudal* gene in its homeodomain sequence and early expression *Mech. Dev.* **43**, 71–78 (1993).
- Beck, F., Erler, T., Russell, A. & James, R. Expression of *Cdx-2* in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes *Dev. Dyn.* **204**, 217–229 (1995).
- James, R., Erler, T. & Kazenwadel, J. Structure of the murine homeobox gene *Cdx-2*: Expression in embryonic and adult intestinal epithelium *J. Biol. Chem.* **269**, 15229–15237 (1994).
- Capecchi, M. R. Altering the genome by homologous recombination *Science* **244**, 1288–1292 (1989).
- Sabramanian, V., Meyer, B. & Gruss, P. Disruption of the murine homeobox gene *Cdx1* affects axial skeletal identities by altering the mesodermal expression of *Hox* genes *Cell* **83**, 641–653 (1995).
- Ee, H. C., Erler, T., Bhatnagar, P. S., Young, G. P. & James, R. *Cdx-2* homeodomain protein expression in human and rat colorectal adenoma and carcinoma *Am. J. Pathol.* **147**, 586–592 (1995).
- Slack, J. M. W. Homeotic transformations in man: implications for the mechanism of embryonic development and for the organisation of epithelia *J. Theor. Biol.* **114**, 463–490 (1985).
- Rubin, H. Cancer as a dynamic developmental disorder *Cancer Res.* **45**, 2935–2942 (1985).
- Moser, A. R., Rapaich, A., Pitot, H. & Dove, W. E. A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse *Science* **247**, 322–324 (1990).
- Moser, A. R. *et al.* Homozygosity for the *Min* allele of *Apc* results in disruption of mouse development prior to gastrulation *Dev. Dyn.* **203**, 422–433 (1995).
- Chawengsaksophak, K. & Beck, F. Chromosomal localisation of *Cdx2*, a murine homologue of the *Drosophila* gene *caudal* to mouse chromosome 5 *Genomics* **34**, 270–271 (1996).
- Luongo, C. *et al.* Mapping of multiple intestinal neoplasia (*Min*) to proximal chromosome 18 of the mouse *Genomics* **15**, 3–8 (1993).
- Duboule, D. Vertebrate *Hox* genes and proliferation: an alternative pathway to homeosis *Curr. Opin. Genet. Dev.* **5**, 525–528 (1995).
- German, M. S. *et al.* Localisation of the genes encoding two transcription factors, *LMX1* and *CDX3* regulating insulin gene expression to human chromosomes 1 and 13 *Genomics* **24**, 403–404 (1994).
- Tybuliewicz, V. L. *et al.* Neonatal lethality and lymphopenia in mice with a homozygous disruption of the *c-abl* proto-oncogene *Cell* **65**, 1153–1163 (1991).
- Szabo, P. & Mann, J. R. Expression and methylation of imprinted genes during *in vitro* differentiation of mouse parthenogenetic and androgenetic embryonic stem cell lines *Development* **120**, 1651–1660 (1994).
- Köntgen, F. & Stewart, C. L. Simple screening procedure to detect gene targeting events in embryonic stem cells. In *Methods in Enzymology* (eds Wassarman, P. M. & De Pamphilis, M. L.) 878–890 (Academic, San Diego, California, 1993).
- Robb, L. *et al.* Absence of yolk sac haematopoiesis from mice with a targeted disruption of the *scl* gene *Proc. Natl Acad. Sci. USA* **92**, 7075–7079 (1995).
- Ramirez-Solis, R., Davis, A. C. & Bradley, A. Gene targeting in embryonic stem cells. In *Methods in Enzymology* (eds Wassarman, P. M. & De Pamphilis, M. L.) 855–878 (Academic, San Diego, California, 1993).
- Inouye, M. Differential staining of cartilage and bone in fetal mouse skeleton by Alcian Blue and Alizarin Red S *Congen. Anom.* **16**, 171–173 (1976).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank C. Andrews, N. Corlett, H. Valcanis and J. Tucci for technical assistance; D. Bowtell for advice at the inception of this project; T. Epp, V. Harley and P. Senior for discussion; and S. Gabriel for donating materials. This work was supported by an NH&MRC block grant to the Howard Florey Institute. R.J. was supported by an R.D. Wright fellowship from NH&MRC.

Correspondence should be addressed to F.B. (e-mail: f.beck@howard.florey.unimelb.edu.au). Requests for materials should be addressed to R.J. (e-mail: rjames@medicine.adelaide.edu.au) or K.C. (e-mail: kallayane@chawengsaksophak.hfist@muwaye.unimelb.edu.au).

Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy

B. Böttcher*, S. A. Wynne & R. A. Crowther

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Hepatitis B virus, a major human pathogen with an estimated 300 million carriers worldwide, can lead to cirrhosis and liver cancer in cases of chronic infection. The virus consists of an inner nucleocapsid or core, surrounded by a lipid envelope containing virally encoded surface proteins. The core protein, when expressed in bacteria, assembles into core shell particles, closely resembling the native core of the virus. Here we use electron cryomicroscopy to solve the structure of the core protein to 7.4 Å resolution. Images of about 6,400 individual particles from 34 micrographs at different levels of defocus were combined, imposing icosahedral symmetry. The three-dimensional map reveals the complete fold of the polypeptide chain, which is quite unlike previously solved viral capsid proteins and is largely α -helical. The dimer clustering of subunits produces spikes on the surface of the shell, which consist of radial bundles of four long α -helices. Our model implies that the sequence corresponding to the immunodominant region of the core protein lies at the tip of the spike and also explains other properties of the core protein.

Hepatitis B virus core protein (HBc), when expressed in *Escherichia coli*¹, assembles into spherical shells resembling those seen in infected liver². Icosahedral shells of two sizes are produced, containing 180 ($T = 3$, where T is the triangulation number and the shell contains $60T$ subunits or 240 ($T = 4$) subunits³. The subunits are clustered as dimers, producing spikes that stick out from the underlying shell domain. The overall diameters of the two kinds of shell are 320 and 360 Å across the tips of the spikes. The core protein is 183 amino acids long and contains a very basic carboxy-terminal region that is believed to interact with nucleic acid. When this is removed, provided the truncation is not before amino acid 140, the core protein still assembles into shells but no longer packages nucleic acid⁴. The precise truncation point influences the distribution between $T = 3$ and $T = 4$ shells, with longer proteins favouring the $T = 4$ configuration⁵. For the protein used here, truncated after residue 149, about 98% of the shells are of the $T = 4$ form, as judged by electron microscopy. Circular dichroism measurements have indicated the presence of substantial amounts of α -helix⁶, amounting to 65% for the C-terminally truncated protein⁷.

The three-dimensional map of the $T = 4$ shell is shown in Fig. 1, first as a view of the whole shell and second as a magnified view of part of the structure. The map was computed from images of 6,384 individual particles taken from 34 micrographs. The large number of views was needed because the signal-to-noise ratio for the unstained specimen is poor in the higher-resolution ranges. The resolution attained at each stage of processing and in the final map was assessed by Fourier shell correlations (Fig. 2). The correlation coefficient measures the agreement in each resolution band between two independent three-dimensional maps computed by splitting in half the data set created at each stage. For the final icosahedrally averaged map, the correlation coefficient drops to 0.5 at $1/7.4 \text{ Å}^{-1}$, indicating the effective resolution of the map.

The $T = 4$ shell contains 240 subunits clustered in 120 dimers (Fig. 1a). Although icosahedral symmetry has been imposed, the

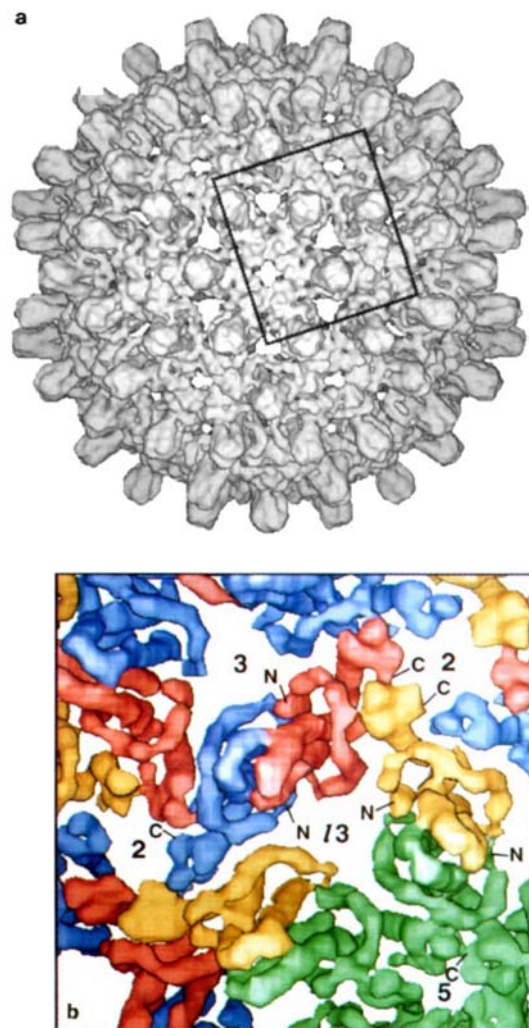


Figure 1 Three-dimensional map of the hepatitis B virus core protein shell, represented as a shaded surface. **a**, The whole shell containing 240 subunits arranged as 120 dimers, viewed down a strict 2-fold axis. **b**, Close-up of part of the shell indicated by the rectangular box in **a**, viewed down a local 3-fold axis (indicated by /3). The positions of neighbouring strict 2-fold, 3-fold and 5-fold axes are marked. The density has been coloured to indicate the four symmetrically independent monomers. The positions of the putative N and C termini of the polypeptide are indicated. The threshold for the display surface is higher in **b** than in **a** to reveal the path of the polypeptide more clearly.

map contains four totally independent images of the core subunit, making two independent dimers. The density appears as a series of folded tubular connected regions, of thickness $\sim 6\text{--}9 \text{ Å}$ and separated by $\sim 10\text{--}12 \text{ Å}$ (Fig. 1b). The boundaries of each dimer unit are clearly visible in all regions, except where the dimers meet around the 5-fold and local 6-fold (strict 2-fold in $T = 4$) axes. The two monomers forming each dimer make an extensive interface across each local 2-fold axis. It is possible to assign all the density in the map in a consistent way to one or other of the four independent monomers, such that the strict and local symmetry axes are obeyed (indicated by the colouring in Fig. 1b). The resulting four monomer units all have an almost identical fold. One of the dimers (the red-blue dimer in Fig. 1b) is shown cut out and with the two monomer chains traced in Fig. 3. The chain traces were made with the 'bones' representation in the O display program⁸.

A diagram of the fold of the monomer subunit, viewed from the same direction as the blue chain in Fig. 3a, is shown in Fig. 4. Two

* Present address: Institut für Physikalische Chemie, Universität Freiburg, Albertstrasse 23a, D-79104 Freiburg, Germany.

Figure 2 Fourier shell correlations showing the progress of the structure determination as more particles were included. Traces relate as follows to the approximate number of particles: (1) 200 (1 micrograph); (2) 1,000 (5 micrographs); (3) 3,000 (17 micrographs); (4) 6,384 (34 micrographs); (5) as (4), but with full icosahedral symmetry imposed on the map. In each case the set of particles was divided in half, two independent three-dimensional maps computed and the Fourier shell correlations then calculated. In (1), no correction was made for the contrast transfer function, which manifests itself as ripples that are still seen weakly in (2) after correction and combination of data from more micrographs.

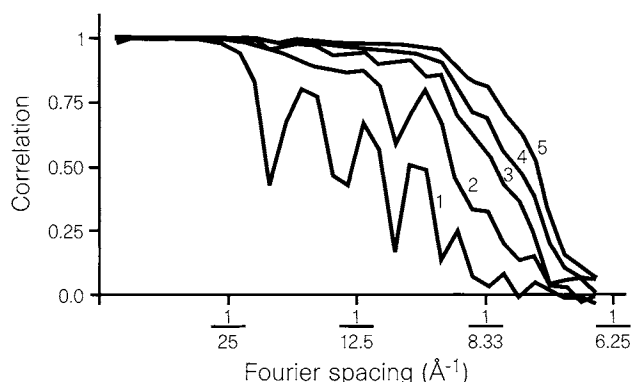
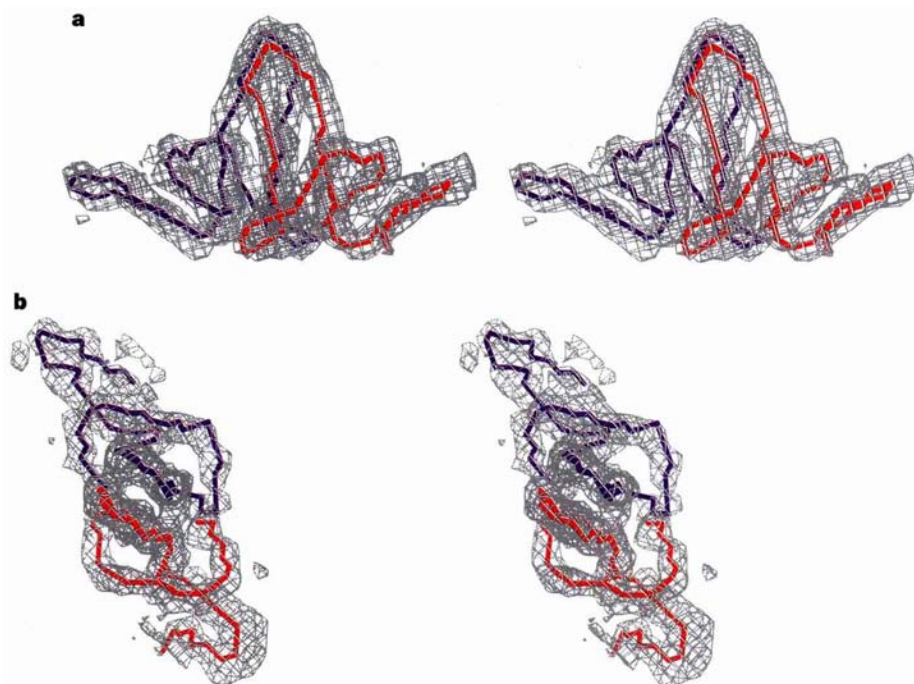


Figure 3 Stereo views of one dimer, the red-blue dimer from Fig. 1b, with the chain traces superimposed. In **a**, the view is roughly normal to the spike, whereas in **b** the view is down the spike from the outside.



tubes of density, each of length about 42 Å, run from the inside of the shell to the tip of the spike, where they are joined by a loop. This main feature of the fold represents, we believe, an antiparallel α -helical hairpin. Each dimer spike thus consists of a four-helix bundle, as suggested on the basis of the dimensions of the spike⁵. The tip of the spike lies roughly half-way along the chain trace. It is known that the principal antigenic site of the protein corresponds to amino acids in the region 78–82 (ref. 9), which contain a conserved proline at residue 79. It is therefore likely that this region of the sequence corresponds to the loop at the tip of the hairpin. For reasons that will become apparent, we believe that the amino terminus of the chain lies in the position indicated in Fig. 4. From here the chain loops around the lower part of the hairpin, including a straight tubular region ~ 15 Å long that appears to be an α -helix, before joining the base of the hairpin. The first long helix of the hairpin is roughly straight, whereas the second long helix forming the other half of the hairpin has a distinct kink about half-way along. A short loop joins the base of the second kinked helix to another straight tubular region ~ 23 Å long, which again probably represents an α -helix. The chain then folds back parallel to this last helix before disappearing into weak density on the inside of the shell. The total length of α -helices noted ($15 + 42 + 42 + 23$ Å) corresponds to about 81 amino acids, which represents 55% of the truncated chain. This value is in good agreement with circular

dichroism measurements⁷, as other regions of the tubular density could well correspond to further short segments of α -helix.

We have attempted to put approximate amino-acid numbers on to the diagram shown in Fig. 4. Taking the loop at the tip of the hairpin to represent amino acids 78–82 and allowing 1.5 Å per amino acid, the straight helix would be bounded by conserved proline residues 50 and 79. The kinked helix would include residues 82–110 and would finish at glycine 111. This helix would contain a conserved glycine at position 94, which might correspond to the position of the kink. The C-terminal helix would then correspond to residues 112–128, being terminated by conserved proline residue 129.

Hybrid core proteins made by insertion of foreign sequences into the immunodominant region around amino acid 80 can still assemble into regular shells^{10,11}, even with inserts as large as 46 amino acids¹². The distally exposed position of the immunodominant loop at the tip of a stable four-helix bundle explains this capacity. Besides the immunodominant region, the core protein has a further epitope in the region of amino acid 130 (ref. 9), and monoclonal antibodies specific for the region 127–133 label well-preserved shells^{13,14}. In our model, this region lies at the end of the C-terminal helix (Fig. 4), forming a loop that is exposed and makes a small protrusion on the surface of the shell close to the 5-fold and local 6-fold positions (Fig. 1). After the tip of the spike, this

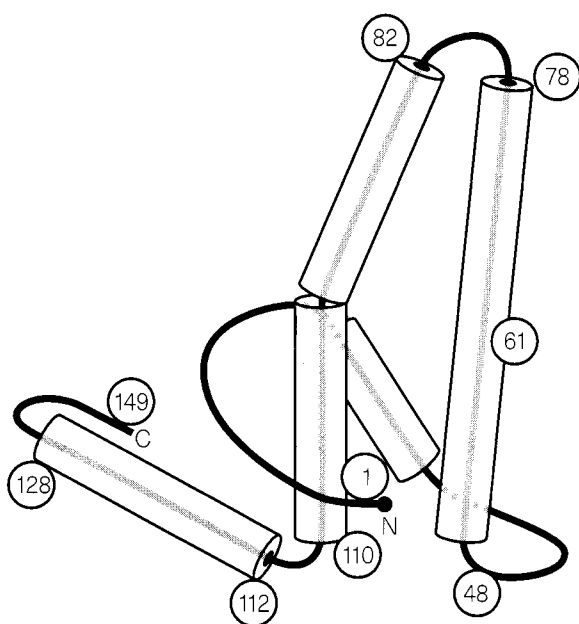


Figure 4 Diagram of the polypeptide fold. The α -helical regions are indicated as cylinders and the putative N and C termini are marked. An approximate numbering scheme for amino acids is given.

protrusion provides the next most exposed and accessible feature on the surface.

Most of the monomer–monomer interface that forms the dimer consists of interactions between the two homologous straight α -helices in the two opposed hairpins. These two helices make an angle of $\sim 20^\circ$. The absolute hand of the map was not determined from the electron microscopy, so we have chosen the hand to make this helix–helix interaction of the more likely $+20^\circ$ type¹⁵. The monomers forming the dimer, on ageing or oxidation of the shells, become linked by homologous intermolecular disulphide bridges, between cysteine 48 and cysteine 48 and between cysteine 61 and cysteine 61 (refs 6, 16). With the numbering scheme shown in Fig. 4, Cys 48 is at the base and Cys 61 about half-way up the straight helix. Both are in positions where they could form disulphide bonds with the homologous residues on the other monomer related by the local 2-fold axis.

The viral genome encodes a secreted monomer form of the core protein, referred to as e-antigen (HBeAg), which in its mature state after cleavage of a signal sequence has a 10-amino-acid extension at the N terminus and Val 149 as the C terminus^{17,18}. The residue at position -7 (relative to the start of the core protein) is a cysteine and in HBeAg this forms a disulphide bond with either Cys 61 or Cys 48^{19,20}. With our assignment of the N terminus, it would be easy to extend the chain backwards in such a way that Cys (-7) could bond with Cys 61, half-way up the straight helix, or with Cys 48, at the base of the straight helix. The formation of a stable loop of this kind across the interfacial surface would block dimerization and therefore shell formation and would ensure that HBeAg remains as a monomer.

The interactions between dimers to form the icosahedral shell occur around the 5-fold and local 6-fold (strict 2-fold in $T = 4$) axes. The principal contacts are mediated by the region of the protein that we ascribe to the C-terminal helix and tail. This part of the protein forms the shell domain at the centres of the rings of 5 and 6 dimers. The quasi-equivalence required to accommodate the different kinds of dimer in the shell appears to be achieved by a tighter and more overlapped packing of the C-terminal segments around the 5-fold axes than occurs around the local 6-fold axes. With this interpretation of the density, it is clear why truncation of

the core protein before residue 140 inhibits shell assembly²¹, as these contacts could not then be formed. These contacts also explain why for C-terminally truncated proteins the precise truncation point is likely to influence the distribution between $T = 3$ and $T = 4$ shells in bacterially expressed protein⁵. There may also be contacts around the strict and local 3-fold axes between residues near the N terminus of one subunit and the N-terminal helix of a subunit in the neighbouring dimer. Contacts of this type might explain why N-terminal deletions appear to inhibit shell assembly²². Finally, the putative position of the C terminus in our map of the truncated core protein is entirely consistent with the continuation of the basic C-terminal tail of the intact protein into the interior of the shell to interact with nucleic acid.

In conclusion, the deduced fold is consistent with a large body of information about the core protein. The fold and resulting shell architecture are quite different from the standard β -sheet jelly-roll packings found in many small viral capsids. Although a spacing of only 7.4 Å has been reached, there is nothing intrinsically limiting the resolution, other than a gradually declining signal-to-noise ratio. There is every reason to believe that inclusion of more images, together with improved corrections for electron optical parameters, should allow calculation of a map at higher, near-atomic resolution. □

Methods

Preparation of core shells. The coding region corresponding to amino acids 1–149 of the full-length HBC protein was recloned by polymerase chain reaction (PCR) from plasmid pHBC3 encoding the full-length core protein from a Latvian isolate of hepatitis B virus²³. Primers were designed to introduce an *NdeI* site at the ATG initiation codon and a stop codon plus *EcoRI* site after residue 149. This C-terminally truncated protein (HBC Δ) lacks the basic region believed to interact with nucleic acid and assembles into empty shells predominantly of the $T = 4$ (240 subunit) type³. The PCR product cleaved with *NdeI* and *EcoRI* was ligated into the vector pT7-SC, also cut with *NdeI* and *EcoRI*. The plasmid was transformed into BL21 (DE3) cells, which were grown to an absorbance of 0.8–1.0 at 600 nm, induced with IPTG and collected 3 hours later.

Cells were collected by centrifugation and lysed by freezing and thawing ($\times 3$) in lysis buffer (50 mM Tris-HCl, pH 8.0, 5 mM EDTA, 100 $\mu\text{g ml}^{-1}$ PMSF, 2 mg ml^{-1} lysozyme), using 20 ml lysis buffer per litre of cell culture. MgCl_2 (100 mM) and DNase (200 $\mu\text{g ml}^{-1}$) were added (2 ml per litre of cell culture). Cell debris was removed by centrifugation (at 15,000 r.p.m. for 30 min). To the supernatant, $(\text{NH}_4)_2\text{SO}_4$ was added to 30% and the precipitate was pelleted at 15,000 r.p.m. for 10 min. The pellet was resuspended in Tris-buffered saline (TBS) containing 0.1% Triton X-100 and applied to a Sepharose–CL4B column. Fractions containing core protein, detected by SDS–PAGE, were pooled and precipitated with 50% $(\text{NH}_4)_2\text{SO}_4$. The precipitate was taken up in TBS, dialysed for 4 hours, loaded onto a 10–40% sucrose gradient and spun in an SW28Ti rotor at 28,000 r.p.m. for 17 h. Fractions from the leading edge of the peak of core protein were pooled, dialysed and concentrated to 10 mg ml^{-1} protein.

Electron microscopy. Samples were frozen as described²⁴. 2 μl of 10 mg ml^{-1} HBC Δ suspension were applied to a perforated carbon-coated copper/rhodium grid mounted in a modified controlled environment freezing apparatus²⁵. Then samples were blotted with a double layer of filter paper (Whatman no. 1) for 15–17 s before plunging into liquid ethane.

For electron cryomicroscopy, a Hitachi HF2000 microscope equipped with a field emission gun and a Gatan cold stage was used. Micrographs were taken at a nominal magnification of 60,000 $\times$ with a beam diameter of 1.3–1.5 μm and exposure times between 1.5 and 2 s on Kodak SO163 film. The electron dose was between 10–16 e per Å². Care was taken to get a good coverage of the defocus range between 1 and 2 μm . The magnification of the microscope was calibrated with a vitrified sample of tobacco mosaic virus.

Image processing. Micrographs showing a crisp appearance of the core particles over the holes and a regular granularity of the carbon film were checked by optical diffraction for the absence of astigmatism and for visibility of Thon rings to 1/9 Å⁻¹ in all directions. Selected micrographs were scanned

with a Zeiss-SCAI scanner using a step size of 14 μm per pixel, corresponding to 2.3 $\text{\AA}$ on the specimen. In total 34 micrographs were scanned, from which 6,630 particles were selected interactively using Ximdisp²⁶. From the coordinate files created, particles were then automatically boxed, floated and normalized in their grey-value distribution. Particles from each micrograph were processed as a group.

Orientations and origins of particles were determined and refined using as a reference the current best three-dimensional map, which was improved stepwise by inclusion of particles from additional micrographs and by rounds of parameter refinement. The previously published map at 30 $\text{\AA}$ resolution was taken as the starting point³. The defocus value of each micrograph was determined initially from the positions of the zeros in the sum of squares of transforms of the particles. Later the defocus values were refined by maximizing the Fourier correlation between the current best map after contrast transfer function (c.t.f.) correction and the corrected individual maps from each micrograph. A set of six different projections of the current best map, modified by the c.t.f. appropriate for the defocus of the micrograph being analysed, was used as reference.

Cross common lines between the transforms of this set of projections and the transform of the raw particle were computed to determine the orientation, origin and relative scaling. The initial refinement was performed in a two-step process. In the first step, a complete asymmetric unit search in 2° steps was performed, including information between 1/66 and 1/23 $\text{\AA}^{-1}$. Using the orientation found, the best origin was then determined. In the second step, a search limited in angular range ($\pm 2^\circ$ around the previously determined orientation) was carried out in 0.5° steps for particles with a cross common lines phase residual lower than 85° up to 1/22 $\text{\AA}^{-1}$ (criterion for inclusion in the data set). For this step, information was used between 1/30 $\text{\AA}^{-1}$ and the Fourier spacing where the phase residual of the current best map reached 45°. Then for the resulting best orientation and origin, the radial scaling to the best map was determined. Individual three-dimensional maps with 52 symmetry²⁷ were calculated for each micrograph from particles which fulfilled the inclusion criterion. The resulting maps from the different micrographs were then combined and c.t.f.-corrected as described²⁸. The second step of refinement was repeated with the current best map until there was no further improvement. Eventually 6,384 particles, corresponding to 96% of those originally selected, were included.

The reliability and resolution of the combined three-dimensional maps were assessed at each stage by dividing the particles into two sets and computing independent maps, whose agreement was measured by Fourier shell correlation and phase residual²⁹. For the final map, the phase residual rose to 42° as the correlation coefficient fell to 0.5 at a spacing of 7.4 $\text{\AA}$. The map discussed here was icosahedrally averaged and sharpened by applying an inverse temperature factor of 500 $\text{\AA}^2$. Maps were displayed either as shaded surface representations³⁰ or by using the crystallographic program O³.

14. Kenney, J. M., Bonsdorff, C.-H.v., Nassal, M. & Fuller, S. D. Evolutionary conservation in the hepatitis B virus core structure: comparison of human and duck cores. *Structure* **3**, 1009–1019 (1995).
15. Chothia, C. & Finkelstein, A. V. The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007–1039 (1990).
16. Nassal, M., Reiger, A. & Steinau, O. Topological analysis of the hepatitis B virus core particle by cysteine–cysteine cross-linking. *J. Mol. Biol.* **225**, 1013–1025 (1992).
17. Takahashi, K. *et al.* Immunochemical structure of hepatitis B antigen in the serum. *J. Immunol.* **130**, 2903–2907 (1983).
18. Takahashi, K. *et al.* Molecular heterogeneity of e antigen polypeptides in sera from carriers of hepatitis B virus. *J. Immunol.* **147**, 3156–3160 (1991).
19. Wasenauer, G., Köck, J. & Schlicht, H.-J. Relevance of cysteine residues for biosynthesis and antigenicity of human hepatitis B virus e protein. *J. Virol.* **67**, 1315–1321 (1993).
20. Nassal, M. & Reiger, A. An intramolecular disulfide bridge between Cys-7 and Cys61 determines the structure of the secretory core gene product (e antigen) of hepatitis B virus. *J. Virol.* **67**, 4307–4315 (1993).
21. Birnbaum, F. & Nassal, M. Hepatitis B virus nucleocapsid assembly: primary structure requirements in the core protein. *J. Virol.* **64**, 3319–3330 (1990).
22. Khudayakov, U. E. *et al.* The effect of the structure of the terminal regions of the hepatitis B virus gene C polypeptide on the formation of core antigen (HBcAg) particles. *Biomed. Sci.* **2**, 257–265 (1991).
23. Bichko, V., Pushko, P., Dreilina, D., Pumpen, P. & Gren, E. Subtype ayw variant of hepatitis B virus: DNA primary structure analysis. *FEBS Lett.* **185**, 208–212 (1985).
24. Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Quart. Rev. Biophys.* **21**, 129–228 (1988).
25. Bellare, J. R., Davis, H. T., Scriven, L. E. & Talmon, Y. Controlled environment vitrification system: an improved sample preparation technique. *J. Electron Microsc. Tech.* **10**, 87–111 (1988).
26. Crowther, R. A., Henderson, R. & Smith, J. M. MRC image processing programs. *J. Struct. Biol.* **116**, 9–16 (1996).
27. Crowther, R. A. Procedures for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs. *Phil. Trans. R. Soc. (Lond.)* **B 261**, 221–230 (1971).
28. Böttcher, B. & Crowther, R. A. Difference imaging reveals ordered regions of RNA in turnip yellow mosaic virus. *Structure* **4**, 387–394 (1996).
29. Heel, M.v. Similarity measures between images. *Ultramicroscopy* **21**, 95–100 (1987).
30. Vigers, G. P. A. Clathrin Assemblies in Vitro Ice. Thesis, Univ. Cambridge (1986).

Acknowledgements. We thank P. Pumpens for a clone of the core protein, P. J. G. Butler for advice on protein purification, B. Pashley for help with illustrations, and our many colleagues for suggestions and discussions. B.B. was supported by an EC Human Capital and Mobility Grant.

Correspondence and requests for materials should be addressed to R.A.C. (e-mail: rac1@mrc-lmb.cam.ac.uk).

Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy

J. F. Conway*, N. Cheng*, A. Zlotnick*, P. T. Wingfield†, S. J. Stahl† & A. C. Steven*

* Laboratory of Structural Biology and † Protein Expression Laboratory, National Institute of Arthritis, Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

Despite the development of vaccines, the hepatitis B virus remains a major cause of human liver disease¹. The virion consists of a lipoprotein envelope surrounding an icosahedral capsid composed of dimers of a 183-residue protein, 'core antigen' (HBcAg)². Knowledge of its structure is important for the design of antiviral drugs, but it has yet to be determined. Residues 150–183 are known to form a protamine-like domain required for packaging RNA, and residues 1–149 form the 'assembly domain' that polymerizes into capsids² and, unusually for a capsid protein, is highly α -helical³. Density maps calculated from cryo-electron micrographs^{4–6} show that the assembly domain dimer is T-shaped: its stem constitutes the dimer interface and the tips of its arms make the polymerization contacts. By refining the procedures used to calculate the map, we have extended the resolution to 9 $\text{\AA}$, revealing major elements of secondary structure. In particular, the stem, which protrudes as a spike on the capsid's outer surface, is a 4-helix bundle, formed by the pairing of α -helical hairpins from both subunits.

The core antigen (HBcAg) assembles into capsids of two sizes, with icosahedral triangulation numbers of $T = 3$ (~280 $\text{\AA}$ diameter, 90 dimers), and $T = 4$ (310 $\text{\AA}$, 120 dimers), respectively^{3,4}. The most detailed structural information to date has emerged from

Received 21 November 1996; accepted 9 January 1997.

1. Pasek, M. *et al.* Hepatitis B virus genes and their expression in *E. coli*. *Nature* **282**, 575–579 (1979).
2. Cohen, B. J. & Richmond, J. E. Electron microscopy of hepatitis B core antigen synthesized in *E. coli*. *Nature* **296**, 677–678 (1982).
3. Crowther, R. A. *et al.* Three-dimensional structure of hepatitis B virus core particles determined by electron cryomicroscopy. *Cell* **77**, 943–950 (1994).
4. Gallina, A. *et al.* A recombinant hepatitis B core antigen polypeptide with the protamine-like domain deleted self-assembles into capsid particles but fails to bind nucleic acids. *J. Virol.* **63**, 4645–4652 (1989).
5. Zlotnick, A. *et al.* Dimorphism of hepatitis B virus capsids is strongly influenced by the C-terminus of the capsid protein. *Biochemistry* **35**, 7412–7421 (1996).
6. Zheng, J., Schödel, E. & Peterson, D. L. The structure of hepadnaviral core antigens. *J. Biol. Chem.* **267**, 9422–9429 (1992).
7. Wingfield, P. T., Stahl, S. J., Williams, R. W. & Steven, A. C. Hepatitis core antigen produced in *Escherichia coli*: subunit composition, conformational analysis, and *in vitro* capsid assembly. *Biochemistry* **34**, 4919–4932 (1995).
8. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
9. Salfeld, J., Pfaff, E., Noah, M. & Schaller, H. Antigenic determinants and functional domains in core antigen and e antigen from hepatitis B virus. *J. Virol.* **63**, 798–808 (1989).
10. Brown, A. L. *et al.* Foreign epitopes in immunodominant regions of hepatitis B core particles are highly immunogenic and conformationally restricted. *Vaccine* **9**, 595–601 (1991).
11. Schödel, F. *et al.* The position of heterologous epitopes inserted in hepatitis B virus core particles determines their immunogenicity. *J. Virol.* **66**, 106–114 (1992).
12. Borisova, G. *et al.* Hybrid hepatitis B virus nucleocapsid bearing an immunodominant region from hepatitis B virus surface antigen. *J. Virol.* **67**, 3696–3701 (1993).
13. Pushko, P. *et al.* Identification of hepatitis B virus core protein regions exposed or internalized at the surface of HBcAg particles by scanning with monoclonal antibodies. *Virology* **202**, 912–920 (1994).

cryo-electron microscopy at resolutions of ~ 30 Å to ~ 17 Å (refs 4–6). These studies have shown that the capsid is highly fenestrated and the dimer is composed of two L-shaped subunits, back-to-back. We have now been able to double the resolution, mainly by two innovations. First, we improved our procedures for correcting the images for phase-contrast effects. These effects⁷, which restrict the resolution, are incurred upon defocusing the beam in order to generate contrast in images of ice-embedded specimens. Second, we refined our determination of the orientation and alignment para-

eters of each capsid image. These provisions made it possible to include much larger numbers of images in the reconstruction, and thus suppress the high noise levels encountered in the high resolution data.

The $T = 4$ capsid model shown in Fig. 1 was calculated from 600 image pairs. Its resolution was assessed consistently by three criteria: the Fourier ring correlation⁸ (8.4 Å), the Fourier shell correlation⁹ (9.1 Å), and the 'reliability index'¹⁰ (9.0 Å). The $T = 3$ capsid map, based on 245 particles, was noisier and had slightly lower resolution

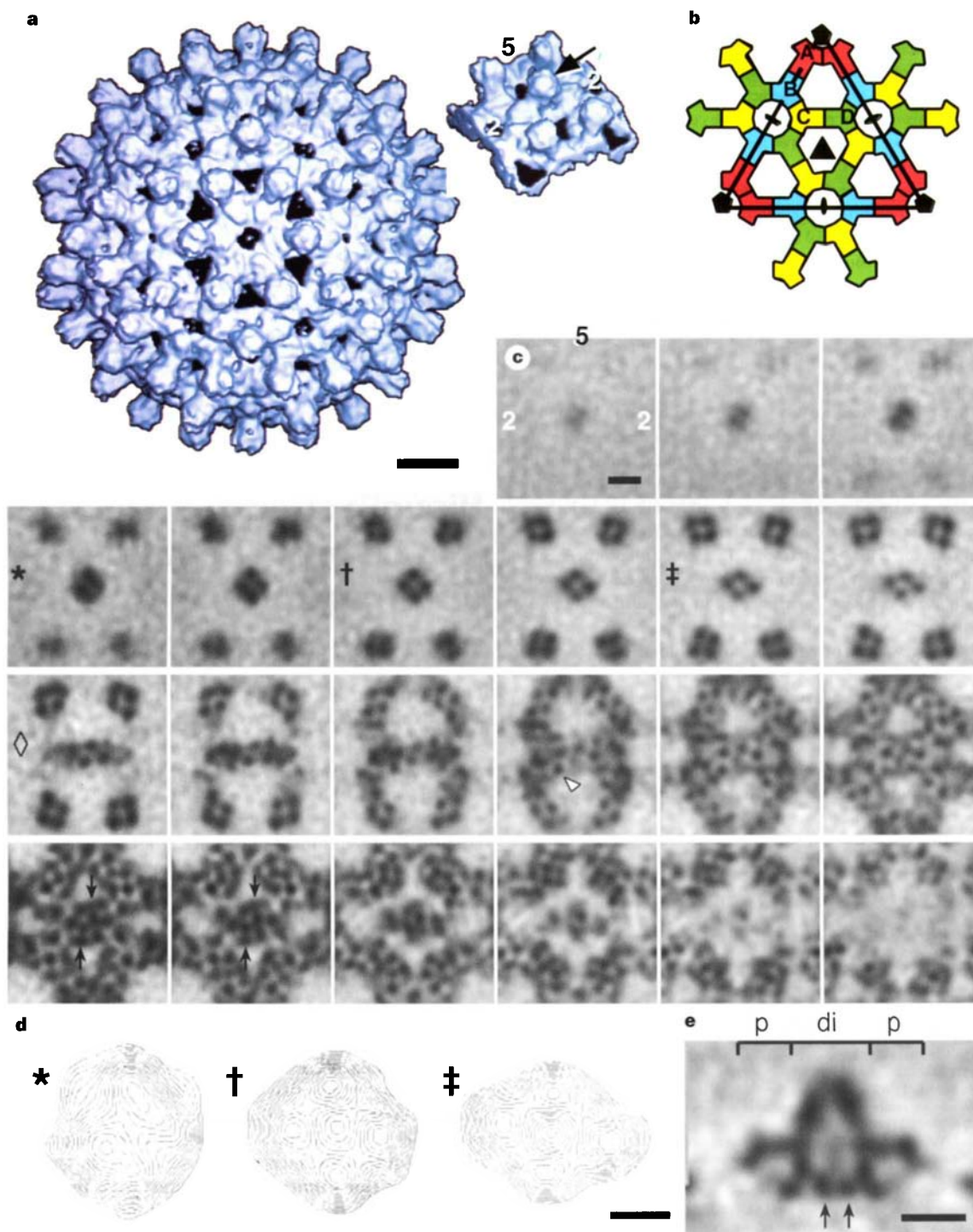


Figure 1 The $T = 4$ capsid of HBV, determined at 9 Å resolution from cryo-electron micrographs. **a**, The capsid exterior, viewed along a 2-fold axis; scale bar, 50 Å. **b**, Diagram showing distribution of subunits (A, B, C, D) and dimers (A-B, C-D) on this lattice. **c**, Serial tangential sections through the portion of the map shown in **a**,

inset, upper right), centred on a C-D dimer. The sections are 1.3 Å thick and spaced 2.6 Å apart; scale bar, 25 Å. **d**, Contour plots of three sections (*, †, ‡); scale bar, 10 Å. **e**, Transverse central section through a C-D dimer; scale bar, 25 Å. pol = polymerization domain, dim = dimerization domain.

(~ 10 Å) but revealed the same features (data not shown). Both maps are notably clean and well defined, as attested by the clear demarcation between protein and solvent (see Fig. 1c).

A surface lattice diagram (Fig. 1b) distinguishes the four kinds of quasi-equivalent subunits (A, B, C, D) and the two dimers (A–B and C–D) present on the $T = 4$ capsid. Inset in Fig. 1a is an area centred on a C–D spike, whose internal structure is shown in serial tangential sections in Fig. 1c. Proceeding inwards from its outer tip, spike density first resolves into two elongated masses, which then separate into four smaller, equally sized, masses centred on the corners of a ~ 10 Å square. Proceeding inwards, this square becomes increasingly rhomboid. The two outer densities (those at the ends of the long axis of the rhombus, Fig. 1c†) extend inwards to the inner surface of the capsid. On leaving the spike, the two inner densities diminish in intensity but they also continue unbroken to the inner surface. Identical substructure was visualized in the A–B spike, and in the A–B and C–C spikes of the $T = 3$ capsid except that the diminution in density in the lower part of the inner column was less pronounced (data not shown). As the two distinct spikes on each capsid were not averaged together in calculating the maps, their consistency attests to a fourfold reproducibility of these results. Entering the contiguous shell, additional features are seen.

The resolution achieved, ~ 9 Å, is sufficient to resolve individual α -helices in parallel bundles in which their centre-to-centre spacing is ~ 10 Å, although it would not allow one to distinguish with confidence between an α -helix and, say, a two-stranded β -sheet. Nevertheless, knowing that HBcAg has a high α -helix content and a dearth of β -sheet³, we favour the interpretation that the four densities resolved in the spike are α -helices. They are depicted in contour plots in Fig. 1d. This interpretation is supported by the observed inter-helix spacing (10 Å), and the crossing angle ($\sim 20^\circ$) being similar to values observed in other 4-helix bundles^{11,12}. The

map (Fig. 1c, panels *, †) shows clearly how the four helices connect into two hairpins. The two outer helices are ~ 40 Å long, about 26 residues. As noted above, the two inner helices appear less dense on entering the contiguous shell, raising the possibility that the polypeptide chain may assume some other extended conformation in this region. We infer that formation of this bundle from two α -helical hairpins, oriented in parallel, is a major factor in dimerization.

These conclusions were based in part on evaluation of the density map by molecular modelling (Fig. 2). Pairs of dyad-related α -helical hairpins fitted well into the density of each spike. Elsewhere, this model is 'domainal', not molecular, and is intended to convey the overall layout of the capsid protein. Nevertheless, in two other sites, we see possible α -helices, tentatively identified as such on the grounds that their cross-sectional size and density are similar to those of the spike helices. One is adjacent and roughly parallel to the bottom segment of the outer spike helix (Fig. 1c, arrowhead) and it connects to the other putative helix, a short one that lies tangentially to and on the inner surface (Fig. 1c, arrows). Two copies of the latter motif may interact across the dimer interface to stabilize its base.

Spectroscopic data indicate that 80–100 residues are α -helical³. The hairpin accounts for 40–50 residues with this conformation, and the two shorter candidate helices for another 15–20. Although it is premature to embark on tracing the polypeptide chain, the outer spike helix does provide a reference point, and it has been suggested¹³ that a helix of about this length, from amino acid 95 to 122, is involved in dimerization. This sequence is followed by an extended proline-rich region (residues 128 to 144), which presumably is not α -helical. Because the loop at the top of the spike must be short, the putative outer helix (residues 95–122) should be oriented with its amino terminus at the top of the spike. Proximally, there would be the hairpin loop (~ 5 residues), preceded by the other

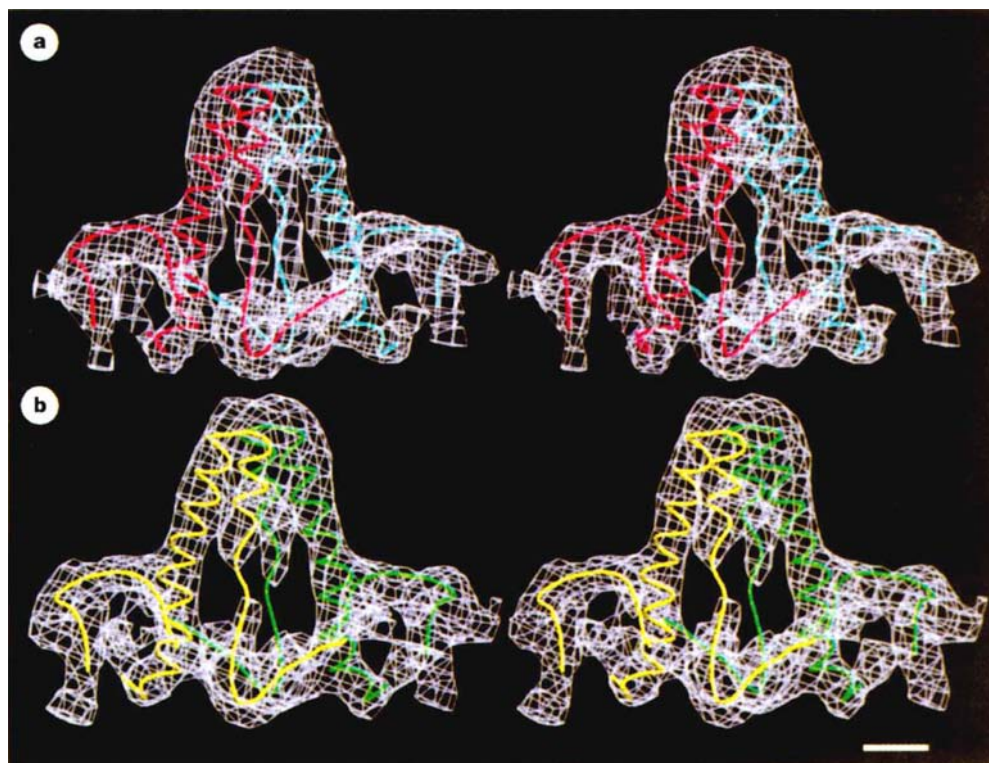


Figure 2 Stereo pairs of density maps (rendered as meshed envelopes by program O³⁰, contoured at 2.75σ), with 'wire' models of HBcAg dimers, viewed tangentially to the capsid surface (see Fig. 1e). The A subunit is red, B blue, C yellow, D green (see Fig. 1b); scale bar, 10 Å. A close similarity, not imposed by icosahedral averaging, was observed between the two kinds of dimers and

between both halves of each dimer. The handedness of the map has yet to be determined, and the enantiomer shown here was chosen arbitrarily. However, because the columns of density identified as α -helices (which are known to be right-handed) are rather straight, the alternative enantiomer could be modelled equally well at the current resolution.

helix. This would place the immunodominant epitope, residues 76 to 82¹⁴, on the side of the spike just above the contiguous shell (see Fig. 2). Alternatively, the epitope has been proposed to lie on top of the spike⁴. Another recent model, based on sequence-derived secondary structure predictions¹⁵, also envisages the spike to be a 4-helix bundle and assigns this epitope to positions around the 5-fold and 6-fold axes.

Although most viral capsid proteins solved to date exhibit the 'jelly roll' β -barrel fold¹⁶, there have been a few recent reports of other α -helical capsid proteins^{17–19}. Even at our current resolution of 9 Å, it is apparent that the fold of HBcAg must differ from those of the latter two proteins^{18,19}. However, we note that several non-viral proteins also dimerize through formation of intermolecular 4-helix bundles^{20,21}, although there is considerable scope for variation in this seemingly simple theme. In Rop protein²², for example, two helical hairpins mate in antiparallel, not nearly parallel as in HBcAg; and in interleukin-5²³, three of the four helices are contributed by one subunit.

In this study, we have resolved individual α -helices by cryo-electron microscopy of 'single' particles, a result that follows earlier successes with highly ordered two-dimensional crystals²⁴ and helical polymers^{25,26}. A resolution of 9 Å opens up the possibility of visualizing elements of secondary structure and their mutual disposition²⁷, and is particularly promising for proteins whose α -helix content is high, and whose helices are relatively long and pack laterally into bundles. Moreover, these proteins should be assembled into particles that are large enough and have sufficient surface contrast for their various views to be identified with the requisite precision. With respect to HBcAg, we are optimistic that further advances in cryo-electron microscopy and image processing will make it possible to delineate the full tertiary structure of this protein. □

Methods

HBcAg (residues 1 to 147) was expressed in *Escherichia coli*, purified under dissociating conditions, reassembled *in vitro*, and visualized in a Philips CM20-FEG electron microscope at a magnification of 38,000 (ref. 6). The maps presented were calculated from a single focal pair of micrographs of a field containing 1,618 $T = 4$ particles and 422 $T = 3$ particles. The rotational offset between them, as digitized (0.28°), was determined from the slopes of lines connecting corresponding particles at opposite corners of the two micrographs. The first zeros of the contrast transfer function were at spacings of $\sim(17 \text{ Å})^{-1}$ (1st exposure) and $\sim(27 \text{ Å})^{-1}$ (2nd exposure), respectively. The coherent illumination from the field-emission gun helped to sustain the signal to high resolution, despite the sizeable defocus values. Initially, the micrographs were scanned at 3.42 Å per pixel. Processed as described previously³ and below, these data yielded a map at $\sim 11 \text{ Å}$ resolution for the $T = 4$ particle. Re-scanning at 2.63 Å per pixel made it possible to extend the resolution to 9 Å. Particle orientations and origins were determined by multiple cycles of the PFT algorithm²⁸. The phase origins obtained by separate PFT analyses of the two micrographs were used to align the members of each focal pair mutually, before combining them. In correcting for contrast transfer, two sets of combined images were generated: for the first, only the requisite phase reversals were performed; for the second, both phase reversals and amplitude attenuation effects were corrected. The first set was used for refinement of orientations and origins, and the second set for calculating density maps. The corrections were performed as described⁶, with some modifications that will be described in detail elsewhere. Of these, the most significant was an additional deconvolution for an apparently exponential decay in amplitudes beyond $(17 \text{ Å})^{-1}$, which presumably represents the combined effects of partial coherence, modulation transfer by the film emulsion and densitometer, and radiation damage. In the final cycle, orientations were refined to within 0.3°, with the upper Fourier limit set at $(14 \text{ Å})^{-1}$. The correlation coefficients were ranked, and all particles with values above an arbitrarily chosen threshold (600 for the $T = 4$ capsid, 245 for the $T = 3$ capsid) were used to calculate the final density maps. In the former case, this corresponds to a total of 36,000 averaging operations. The icosahedral asymmetric unit²⁹ was well covered by the distribution of particle orientations

in each data set. For both capsids, two independent maps were calculated from half data sets, to assess reproducibility and allow calculation of the resolution^{8–10}. All features reported for the 600-particle $T = 4$ map were also visible, albeit somewhat noisier, in both 300-particle reconstructions. A second focal pair of micrographs was also analysed, with generally similar results but slightly lower resolution.

Molecular modelling was performed using the program O³⁰ to trace the path of connected density through the A subunit, and to fit two α -helices into the half-spike that it contributes to the A–B dimer. A copy of this model was then rotated and translated as appropriate to fit the spike portion of the B subunit; and similarly, to model the C and D subunits. In this context, we stress that the local 2-fold symmetry within both dimers and the close similarity of the A–B and C–D dimers are features of the map, and were not imposed by the icosahedral reconstruction formalism.

Received 20 November 1996; accepted 16 January 1997.

- Hollinger, F. B. Hepatitis B virus. In *Fields Virology* 3rd edn (eds Fields, B. N. et al.) 2738–2308 (Lippincott-Raven, Philadelphia, 1996).
- Nassal, M. & Schaller, H. Hepatitis B virus nucleocapsid assembly. In *Virus Strategies* (eds Doefler, W. & Bohm, P.) 41–75 (VCH, Weinheim, 1993).
- Wingfield, P. T., Stahl, S. J., Williams, R. W. & Steven, A. C. Hepatitis core antigen produced in *Escherichia coli*: subunit composition, conformational analysis, and *in vitro* capsid assembly. *Biochemistry* **34**, 4919–4932 (1995).
- Crowther, R. A. et al. Three-dimensional structure of hepatitis B virus core particles determined by electron cryomicroscopy. *Cell* **77**, 943–950 (1994).
- Kenney, J. M., von Bonsdorff, C. H., Nassal, M. & Fuller, S. D. Evolutionary conservation in the hepatitis B virus core structure: comparison of human and duck cores. *Structure* **3**, 1009–1019 (1995).
- Zlotnick, A. et al. Dimorphism of hepatitis B virus capsids is strongly influenced by the C-terminus of the capsid protein. *Biochemistry* **35**, 7412–7421 (1996).
- Zemlin, F. Expected contribution of the field-emission gun to high-resolution transmission electron microscopy. *Micron* **25**, 223–226 (1994).
- Conway, J. F. et al. The effects of radiation damage on the structure of frozen hydrated HSV-1 capsids. *J. Struct. Biol.* **111**, 222–233 (1993).
- van Heel, M. Similarity measures between images. *Ultramicroscopy* **21**, 95–100 (1987).
- Winkelmann, D. A., Baker, T. S. & Rayment, I. Three-dimensional structure of myosin subfragment-1 from electron microscopy of sectioned crystals. *J. Cell Biol.* **114**, 701–713 (1991).
- Chothia, C. Principles that determine the structure of proteins. *Annu. Rev. Biochem.* **53**, 537–572 (1984).
- Presnell, S. R. & Cohen, F. E. Topological distribution of four- α -helix bundles. *Proc. Natl Acad. Sci. USA* **86**, 6592–6596 (1989).
- Yu, M. S., Miller, R. H., Emerson, S. & Purcell, R. H. A. A hydrophobic heptad repeat of the core protein of woodchuck hepatitis-virus is required for capsid assembly. *J. Virol.* **70**, 7085–7091 (1996).
- Salfeld, J., Pfaff, E., Noah, M. & Schaller, H. Antigenic determinants and functional domains in core antigen and e antigen from hepatitis B virus. *J. Virol.* **63**, 798–808 (1989).
- Bringas, R. Folding and assembly of hepatitis B virus core protein. A new model proposal. *J. Struct. Biol.* (in the press).
- Rossmann, M. G. & Johnson, J. E. Icosahedral RNA virus structure. *Annu. Rev. Biochem.* **58**, 533–573 (1989).
- Bamford, J. K., Bamford, D. H., Li, T. & Thomas, G. J. Jr Structural studies of the enveloped dsRNA bacteriophage phi 6 of *Pseudomonas syringae* by Raman spectroscopy. II. Nucleocapsid structure and thermostability of the virion, nucleocapsid and polymerase complex. *J. Mol. Biol.* **230**, 473–482 (1993).
- Momany, C. et al. Crystal structure of dimeric HIV-1 capsid protein. *Nature Struct. Biol.* **3**, 763–770 (1996).
- Rao, Z. et al. Crystal structure of SIV matrix antigen and implications for virus assembly. *Nature* **378**, 743–747 (1995).
- Arents, G., Burlingame, R. W., Wang, B. C., Love, W. E. & Moudrianakis, E. N. The nucleosomal core histone octamer at 3.1 Å resolution: a tripartite protein assembly and a left-handed superhelix. *Proc. Natl Acad. Sci. USA* **88**, 10148–10152 (1991).
- Xie, X. et al. Structural similarity between TAFs and the heterotetrameric core of the histone octamer. *Nature* **380**, 316–322 (1996).
- Banner, D. W., Kokkinidis, M. & Tsernoglou, D. Structure of the ColE1 rop protein at 1.7 Å resolution. *J. Mol. Biol.* **196**, 657–675 (1987).
- Milburn, M. V. et al. A novel dimer configuration revealed by the crystal structure at 2.4 Å resolution of human interleukin-5. *Nature* **363**, 172–176 (1993).
- Unwin, P. N. & Henderson, R. Molecular structure determination by electron microscopy of unstained crystalline specimens. *J. Mol. Biol.* **94**, 425–440 (1975).
- Jeng, T. W., Crowther, R. A., Stubbs, G. & Chiu, W. Visualization of α -helices in tobacco mosaic virus by cryo-electron microscopy. *J. Mol. Biol.* **205**, 251–257 (1989).
- Morgan, D. G., Owen, C., Melanson, L. A. & DeRosier, D. J. Structure of bacterial flagellar filaments at 11 Å resolution: packing of the α -helices. *J. Mol. Biol.* **249**, 88–110 (1995).
- Booy, F. P. et al. High resolution studies of papillomavirus and herpes simplex virus (Abstract). *Proc. Xth Eur. Congr. Microscopy, Dublin* (1996).
- Baker, T. S. & Cheng, R. H. A model-based approach for determining orientations of biological macromolecules imaged by cryoelectron microscopy. *J. Struct. Biol.* **116**, 120–130 (1996).
- Baker, T. S., Drak, J. & Bina, M. The capsid of small papova viruses contains 72 pentameric capsomeres: direct evidence from cryo-electron-microscopy of simian virus 40. *Biophys. J.* **55**, 243–253 (1989).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for binding protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).

Acknowledgements. We gratefully acknowledge helpful discussions with B. Trus, F. Booy and D. Belnap. T. Baker for software, I. Palmer and J. Kaufman for technical support, and support of A.Z. by a NRC-NIAMS Research Associateship.

Correspondence and requests for materials should be addressed to A.C.S. (e-mail: Alasdair_Steven@nih.gov).